Numerical Analysis - Part II

Anders C. Hansen

Lecture 21

Iterative methods for linear algebraic systems

The conjugate gradient method

Here it is.

(A) For any initial vector $\mathbf{x}^{(0)}$, set $\mathbf{d}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$;

(B) For $k \geq 0$, calculate $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$ and the residual

$$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)}, \quad \text{with}$$

$$\alpha_k := \{\, \mathbf{r}^{(k+1)} \perp \mathbf{d}^{(k)} \} = \frac{(\mathbf{r}^{(k)}, \mathbf{d}^{(k)})}{(A\mathbf{d}^{(k)}, \mathbf{d}^{(k)})}, \quad k \geq 0\,. \tag{1}$$

(C) For the same k, the next conjugate direction is the vector

$$\mathbf{d}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{d}^{(k)}, \quad \text{with}$$

$$\beta_k := \{\, \mathbf{d}^{(k+1)} \perp A\mathbf{d}^{(k)} \} = -\frac{(\mathbf{r}^{(k+1)}, A\mathbf{d}^{(k)})}{(\mathbf{d}^{(k)}, A\mathbf{d}^{(k)})}, \quad k \geq 0\,. \tag{2}$$

Theorem 1 (Properties of CGM)

For every $m \geq 0$, the conjugate gradient method has the following properties.

(1) *The linear space spanned by the residuals $\{\mathbf{r}^{(i)}\}$ is the same as the linear space spanned by the conjugate directions $\{\mathbf{d}^{(i)}\}$ and it coincides with the space spanned by $\{A^i\mathbf{r}^{(0)}\}$:*

$$\operatorname{span}\{\mathbf{r}^{(i)}\}_{i=0}^{m} = \operatorname{span}\{\mathbf{d}^{(i)}\}_{i=0}^{m} = \operatorname{span}\{A^i\mathbf{r}^{(0)}\}_{i=0}^{m}\,.$$

(2) *The residuals satisfy the orthogonality conditions:* $(\mathbf{r}^{(m)}, \mathbf{r}^{(i)}) = (\mathbf{r}^{(m)}, \mathbf{d}^{(i)}) = 0$ *for* $i < m$.

(3) *The directions are conjugate (A-orthogonal):* $(\mathbf{d}^{(m)}, \mathbf{d}^{(i)})_A = (\mathbf{d}^{(m)}, A\mathbf{d}^{(i)}) = 0$ *for* $i < m$.

Proof. We use induction on $m \geq 0$, the assertions being trivial for $m = 0$, since $\mathbf{d}^{(0)} = \mathbf{r}^{(0)}$ and (2)-(3) are void. Therefore, assuming that the assertions are true for some $m = k$, we ask if they remain true when $m = k + 1$.

(1) Formula (2)

$$\mathbf{d}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{d}^{(k)}$$

readily implies that equivalence of the spaces spanned by $(\mathbf{r}^{(i)})_0^k$ and $(\mathbf{d}^{(i)})_0^k$, is preserved when k is increased to $k + 1$. Similarly, from $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)}$ in (1), and from the inductive assumption $\mathbf{r}^{(k)}, \mathbf{d}^{(k)} \in \mathrm{span}\{A^i\mathbf{r}^{(0)}\}_{i=0}^{k}$, it follows that $\mathbf{r}^{(k+1)} \in \mathrm{span}\{A^i\mathbf{r}^{(0)}\}_{i=0}^{k+1}$. To see that $A^{k+1}\mathbf{r}^{(0)} \in \mathrm{span}\{\mathbf{r}^{(i)}\}_{i=0}^{k+1}$, since $\alpha_k \neq 0$, the claim follows by (5) if $\mathbf{d}^{(k)}$ has a non-zero component from $A^k\mathbf{r}^{(0)}$, and if not the claim follows from the induction hypothesis.

Proof. Cont. (2) Turning to assertion (2), we need $\mathbf{r}^{(k+1)} \perp \mathbf{r}^{(i)}$ for $i \leq k$, which by (1) is equivalent to

$$\mathbf{r}^{(k+1)} \perp \mathbf{d}^{(i)} \quad \text{for} \quad i \leq k\,.$$

We have $\mathbf{r}^{(k+1)} \perp \mathbf{d}^{(k)}$ by the definition of α_k in (1), so we need

$$\mathbf{r}^{(k+1)} \overset{(1)}{=} \mathbf{r}^{(k)} - \alpha_k A\mathbf{d}^{(k)} \perp \mathbf{d}^{(i)} \quad \text{for} \quad i < k\,,$$

and this follow from the inductive assumptions $\mathbf{r}^{(k)} \perp \mathbf{d}^{(i)}$ and $A\mathbf{d}^{(k)} \perp \mathbf{d}^{(i)}$.

Proof. Cont. (3) It remains to justify (3), namely that $\mathbf{d}^{(k+1)}$ defined in (2) satisfies

$$\mathbf{d}^{(k+1)} \perp A\mathbf{d}^{(i)} \quad \text{for} \quad i \leq k\,.$$

The value of β_k in (2) is defined to give $\mathbf{d}^{(k+1)} \perp A\mathbf{d}^{(k)}$, so we need

$$\mathbf{d}^{(k+1)} \overset{(2)}{=} \mathbf{r}^{(k+1)} + \beta_k \mathbf{d}^{(k)} \perp A\mathbf{d}^{(i)} \quad \text{for} \quad i < k\,.$$

By the inductive hypothesis $\mathbf{d}^{(k)} \perp A\mathbf{d}^{(i)}$, hence it remains to establish that $\mathbf{r}^{(k+1)} \perp A\mathbf{d}^{(i)}$ for $i < k$. Now, the formula (1) yields $A\mathbf{d}^{(i)} = (\mathbf{r}^{(i)} - \mathbf{r}^{(i+1)})/\alpha_i$, therefore we require the conditions $\mathbf{r}^{(k+1)} \perp (\mathbf{r}^{(i)} - \mathbf{r}^{(i+1)})$ for $i < k$, and they are a consequence of the assertion (2) for $m = k + 1$ obtained previously. □

Termination property

Corollary 2 (A termination property)

If the conjugate gradient method is applied in exact arithmetic, then, for any $\mathbf{x}^{(0)} \in \mathbb{R}^n$*, termination occurs after at most* n *iterations. More precisely, termination occurs after at most* s *iterations, where* $s = \dim \operatorname{span}\{A^i \mathbf{r}_0\}_{i=0}^{n-1}$ *(which can be smaller than* n*).*

Proof. Assertion (2) of Theorem 1 states that residuals $(\mathbf{r}^{(k)})_{k\geq 0}$ form a sequence of mutually orthogonal vectors in $\mathbb{R}^n$, therefore at most n of them can be nonzero. Since they also belong to the space $\operatorname{span}\{A^i\mathbf{r}_0\}_{i=0}^{n-1}$, their number is bounded by the dimension of that space. □

Definition 3 (The Krylov subspaces)

Let A be an $n \times n$ matrix, $\mathbf{v} \in \mathbb{R}^n$ nonzero, and $m \in \mathbb{N}$. The linear space $K_m(A, \mathbf{v}) := \operatorname{span}\{A^i \mathbf{v}\}_{i=0}^{m-1}$ is called the *m-th Krylov subspace* of $\mathbb{R}^n$.

Theorem 4 (Number of iterations in CGM)

Let $A > 0$, and let s be the number of its distinct eigenvalues. Then, for any $\mathbf{v}$,

$$\dim K_m(A, \mathbf{v}) \leq s \quad \forall\, m\,. \tag{3}$$

Hence, for any $A > 0$, the number of iterations of the CGM for solving $A\mathbf{x} = \mathbf{b}$ is bounded by the number of distinct eigenvalues of A.

Proof. Inequality (3) is true not just for positive definite $A > 0$, but for any A with n linearly independent eigenvectors $(\mathbf{u}_i)$. Indeed, in that case one can expand $\mathbf{v} = \sum_{i=1}^{n} a_i \mathbf{u}_i$, and then group together eigenvectors with the same eigenvalues: for each λ_ν we set $\mathbf{w}_\nu = \sum_{k=1}^{m_\nu} a_{i_k} \mathbf{u}_{i_k}$ if $A\mathbf{u}_{i_k} = \lambda_\nu \mathbf{u}_{i_k}$. Then

$$\mathbf{v} = \sum_{\nu=1}^{s} c_\nu \mathbf{w}_\nu \,, \qquad c_\nu \in \{0, 1\} \,,$$

hence $A^i \mathbf{v} = \sum_{\nu=1}^{s} c_\nu \lambda_\nu^i \mathbf{w}_\nu$, thus for any m we get $K_m(A, \mathbf{v}) \subseteq \operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_s\}$, and that proves (3). By Corollary 2, the number of iteration in CGM is bounded by $\dim K_m(A, \mathbf{r}^{(0)})$, hence the final conclusion. □

Remark 5

Theorem 4 shows that, unlike other iterative schemes, the conjugate gradient method is both iterative and direct: each iteration produces a reasonable approximation to the exact solution, and the exact solution itself will be recovered after n iterations at most.

Simplifying the CGM-algorithm

We now simplify and reformulate the CGM-algorithm.
Firstly, we rewrite expressions for the parameters α_k and β_k in (1)-(2) as follows:

$$\alpha_k = \frac{(\mathbf{r}^{(k)}, \mathbf{d}^{(k)})}{(\mathbf{d}^{(k)}, A\mathbf{d}^{(k)})} \overset{(c)}{=} \frac{\|\mathbf{r}^{(k)}\|^2}{(\mathbf{d}^{(k)}, A\mathbf{d}^{(k)})} > 0\,,$$

$$\beta_k = -\frac{(\mathbf{r}^{(k+1)}, A\mathbf{d}^{(k)})}{(\mathbf{d}^{(k)}, A\mathbf{d}^{(k)})} \overset{(a)}{=} -\frac{(\mathbf{r}^{(k+1)}, \mathbf{r}^{(k+1)} - \mathbf{r}^{(k)})}{(\mathbf{d}^{(k)}, \mathbf{r}^{(k+1)} - \mathbf{r}^{(k)})} \overset{(b)}{=} \frac{\|\mathbf{r}^{(k+1)}\|^2}{(\mathbf{d}^{(k)}, \mathbf{r}^{(k)})} \overset{(c)}{=} \frac{\|\mathbf{r}^{(k+1)}\|^2}{\|\mathbf{r}^{(k)}\|^2} > 0\,.$$

Here, for β, we used in (a) the fact that $A\mathbf{d}^{(k)}$ is a multiple of $\mathbf{r}^{(k+1)} - \mathbf{r}^{(k)}$ by (1), and in (b) orthogonality of $\mathbf{r}^{(k+1)}$ to both $\mathbf{r}^{(k)}, \mathbf{d}^{(k)}$ proved in Theorem 1(2). Then, for both β and α, we used in (c) the property $(\mathbf{d}^{(k)}, \mathbf{r}^{(k)}) = \|\mathbf{r}^{(k)}\|^2$ which follows from (2) with index $k+1$, taking in account orthogonality $\mathbf{r}^{(k+1)} \perp \mathbf{d}^{(k)}$.
Secondly, we let $\mathbf{x}^{(0)}$ be the zero vector.

Standard form of the conjugate gradient method

Here it is.

(1) Set $k = 0$, $\mathbf{x}^{(0)} = 0$, $\mathbf{r}^{(0)} = \mathbf{b}$, and $\mathbf{d}^{(0)} = \mathbf{r}^{(0)}$;

(2) Calculate the matrix-vector product $\mathbf{v}^{(k)} = A\mathbf{d}^{(k)}$ and $\alpha_k = \|\mathbf{r}^{(k)}\|^2/(\mathbf{d}^{(k)}, \mathbf{v}^{(k)}) > 0$;

(3) Apply the formulae $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$ and $\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} - \alpha_k \mathbf{v}^{(k)}$;

(4) Stop if $\|\mathbf{r}^{(k+1)}\|$ is acceptably small;

(5) Set $\mathbf{d}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{d}^{(k)}$, where $\beta_k = \|\mathbf{r}^{(k+1)}\|^2/\|\mathbf{r}^{(k)}\|^2 > 0$;

(6) Increase $k \to k + 1$ and go back to (2).

Standard form of the conjugate gradient method

The total work is dominated by the number of iterations, multiplied by the time it takes to compute $\mathbf{v}^{(k)} = A\mathbf{d}^{(k)}$. Thus the conjugate gradient algorithm is highly suitable when most of the elements of A are zero, i.e. when A is *sparse*.

Preconditioning

In $A\mathbf{x} = \mathbf{b}$, we change variables, $\mathbf{x} = P^T\widehat{\mathbf{x}}$, where P is a nonsingular $n \times n$ matrix, and multiply both sides with P. Thus, instead of $A\mathbf{x} = \mathbf{b}$, we are solving the linear system

$$PAP^T\widehat{\mathbf{x}} = P\mathbf{b} \quad \Leftrightarrow \quad \widehat{A}\widehat{x} = \widehat{\mathbf{b}}\,. \tag{4}$$

Note that symmetry and positive definiteness of A imply that $\widehat{A} = PAP^T$ is also symmetric and positive definite since $(\widehat{A}\mathbf{y}, \mathbf{y}) = (PAP^T\mathbf{y}, \mathbf{y}) = (AP^T\mathbf{y}, P^T\mathbf{y}) > 0$. Therefore, we can apply conjugate gradients to the new system. This results in the solution $\widehat{\mathbf{x}}$, hence $\mathbf{x} = P^T\widehat{\mathbf{x}}$. This procedure is called the *preconditioned conjugate gradient method* and the matrix P is called the *preconditioner*.

Condition number and convergence rate of CGM

The *condition number* of a matrix A is the value $\kappa(A) := \|A\| \cdot \|A^{-1}\|$, so for a symmetric positive definite matrix A it is the ratio between its largest and smallest eigenvalues,

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \geq 1\,.$$

The closer this number is to 1, the faster the convergence is of CGM. More precisely, for the rate of convergnce of CGM, we have the upper estimate

$$\|\mathbf{e}^{(k)}\|_A \leq 2\rho^k \, \|\mathbf{e}^{(0)}\|_A\,, \qquad \rho = \rho_A = \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} < 1\,. \tag{5}$$

The main idea of preconditioning is to pick P in (4) so that $\kappa(\widehat{A})$ is much smaller than $\kappa(A)$, thus accelerating convergence.

Preconditioning – Choosing P

To this end, we note that the similarity transform $B \to C^{-1}BC$ preserves spectrum, hence

$$\kappa(\widehat{A}) = \kappa(PAP^T) = \kappa(P^{-1}[PAP^T]P) = \kappa(AP^TP)\,,$$

and if we set

$$S^{-1} := P^TP =: (QQ^T)^{-1},$$

then it is suggestive to choose S as an approximation to A which is easy to Cholesky-factorize, i.e., $S = QQ^T$ (or already in this form), and then take $P = Q^{-1}$. Then $AP^TP = AS^{-1}$ is close to identity, hence

$$\kappa(\widehat{A}) = \kappa(AP^TP) \approx \kappa(I) = 1 \quad \Rightarrow \quad \kappa(\widehat{A}) \ll \kappa(A)\,,$$

and the preconditioned system (4) will be solved much faster because of (5).

Preconditioning – Extra cost

Each step in the CGM for solving $A\mathbf{x} = \mathbf{b}$ requires one matrix-vector product $A\mathbf{y}$, so with $P = Q^{-1}$, additional expense in each step of the CGM for the preconditioned system (4) while computing $\widehat{A}\mathbf{y} = PAP^T\mathbf{y}$ is two additional computations

$$\mathbf{u} = P^T\mathbf{y} = Q^{-T}\mathbf{y}, \qquad \mathbf{v} = P\mathbf{z} = Q^{-1}\mathbf{z},$$

for some $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, but note that computing $Q^{-1}\mathbf{z}$ is the same as solving the linear system $Q\mathbf{v} = \mathbf{z}$, which is cheap (via forward substitution) as Q is a lower triangular matrix.

Example 6

1) The simplest choice of S is $D = \operatorname{diag} A$, then $P = D^{-1/2}$ in (4).
2) Another possibility is to choose S as a band matrix with small bandwidth. For example, solving the Poisson equation with the five-point formula, we may take S to be the tridiagonal part of A.
3) One can also take $P = L^{-1}$, where L is the lower triangular part of A (maybe imposing some changes). For example, for the Poisson equation, with $m = 20$ hence dealing with 400×400 system, we take P^{-1} as the lower triangular part of A, but change the diagonal elements from 4 to $\frac{5}{2}$. Then we get a computer precision after just 30 iterations.

Preconditioning – Examples

For the tridiagonal system $A\mathbf{x} = \mathbf{b}$ below, we choose the preconditioner as follows.

$$A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix},$$

$$S = QQ^T = \begin{bmatrix} 1 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}.$$

Preconditioning – Examples

The matrix S coincides with A except at the $(1,1)$-entry. The matrix

$$\widehat{A} = Q^{-1} A Q^{-T}$$

for the preconditioned CGM has just two distinct eigenvalues, and we recover the exact solution just in two steps. To see the latter, note that $\widehat{A}$ is similar to

$$Q^{-T} Q^{-1} A = S^{-1} A,$$

hence it has the same spectrum. Since $A = S + \mathbf{e}_1 \mathbf{e}_1^T$, we have

$$S^{-1} A = I + \mathbf{u} \mathbf{e}_1^T \,,$$

a rank-1 perturbation of the identity matrix, with all eigenvalues but one equal 1 (the remaining one equal $1 + u_1$).

Rate of convergence of CGM

Theorem 7

Consider the CGM. We then have the upper estimate

$$\|\mathbf{e}^{(k)}\|_A \leq 2\rho^k \|\mathbf{e}^{(0)}\|_A\,, \qquad \rho = \rho_A = \frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} < 1\,,$$

where $\mathbf{e}^{(k)} = x^* - x^{(k)}$ *and* $x^{(k)}$ *is the k-th output of the CGM.*

Theorem 8 (Non-examinable)

Given $A \in \mathbb{R}^{n\times n}$, $A > 0$, *let* $\{\mathbf{d}^{(k)}\}_{k=0}^{m-1}$ *be a set of the conjugate directions, i.e.,* $(A\mathbf{d}^{(k)}, \mathbf{d}^{(i)}) = 0$ *for* $i < k$, *and consider*

$$F(\mathbf{x}^{(k)}) := \|\mathbf{x}^* - \mathbf{x}^{(k)}\|_A^2 = \|\mathbf{e}^{(k)}\|_A^2 .$$

Then the value of $F(\mathbf{x}^{(m+1)})$ *obtained through the CGM coincides with the minimum of* $F(\mathbf{y})$ *taken over all* $\mathbf{y} = \mathbf{x}^{(0)} + \sum_{k=0}^{m} c_k \mathbf{d}^{(k)}$ *simultaneously, namely*

$$\arg \min_{c_0,\ldots,c_m} F(\mathbf{y}) = \mathbf{x}^{(m+1)} = \mathbf{x}^{(0)} + \sum_{k=0}^{m} \alpha_k \mathbf{d}^{(k)} .$$

Rate of convergence of CGM

Proof of Theorem 7. As we have seen, every direction $\mathbf{d}^{(i)}$ in CGM is a linear combination of the vectors $(A^s\mathbf{r}^{(0)})_{s=0}^{i}$, therefore, any vector of the form $\widehat{\mathbf{x}}^{(k)} = \mathbf{x}^{(0)} + \sum_{i=0}^{k-1} a_i\mathbf{d}^{(i)}$ can be represented as

$$\widehat{\mathbf{x}}^{(k)} = \mathbf{x}^{(0)} + \sum_{i=0}^{k-1} c_i A^i\mathbf{r}^{(0)}\,. \tag{6}$$

Subtracting both parts of (6) from the exact solution $\mathbf{x}^*$ we obtain $\widehat{\mathbf{e}}^{(k)} = \mathbf{e}^{(0)} - \sum_{i=0}^{k-1} c_i A^i\mathbf{r}^{(0)}$, and since $\mathbf{r}^{(0)} = A\mathbf{e}^{(0)}$, we can express the error $\widehat{\mathbf{e}}^{(k)} = \mathbf{x}^* - \widehat{\mathbf{x}}^{(k)}$ as

$$\widehat{\mathbf{e}}^{(k)} = (I - \sum_{i=1}^{k} c_i A^i)\,\mathbf{e}^{(0)} = P_k(A)\,\mathbf{e}^{(0)}, \tag{7}$$

where P_k is a polynomial of degree $\leq k$, which satisfies $P_k(0) = 1$.

Rate of convergence of CGM

Proof. Cont. Now recall from Theorem 8 that, at the k-th stage, the CGM produces the vector $\mathbf{x}^{(k)}$ that minimizes the functional

$$F(\widehat{\mathbf{x}}^{(k)}) = \|\widehat{\mathbf{e}}^{(k)}\|_A^2 = (A\widehat{\mathbf{e}}^{(k)}, \widehat{\mathbf{e}}^{(k)})$$

over all vectors $\widehat{\mathbf{x}}^{(k)}$ of the form $\widehat{\mathbf{x}}^{(k)} = \mathbf{x}^{(0)} + \sum_{i=0}^{k-1} a_i \mathbf{d}^{(i)}$, hence over all $\widehat{\mathbf{e}}^{(k)}$ of the form (7). Expressing $\mathbf{e}^{(0)}$ as $\mathbf{e}^{(0)} = \sum \gamma_i \mathbf{w}_i$, where $(\mathbf{w}_i)$ are orthonormal eigenvectors of A, we find from (7) that $\widehat{\mathbf{e}}^{(k)} = \sum_i \gamma_i P_k(\lambda_i)\mathbf{w}_i$, and $A\widehat{\mathbf{e}}^{(k)} = \sum_i \gamma_i P_k(\lambda_i)\lambda_i \mathbf{w}_i$, and respectively

$$\|\widehat{\mathbf{e}}^{(k)}\|_A^2 = \sum_i [P_k(\lambda_i)]^2 \lambda_i \gamma_i^2 \leq \max_{\lambda \in \sigma(A)} [P_k(\lambda)]^2 \, \|\mathbf{e}^{(0)}\|_A^2 \,.$$

Hence, because of the minimization property of CGM,

$$\|\mathbf{e}^{(k)}\|_A = \min_{P_k} \|\widehat{\mathbf{e}}^{(k)}\|_A \leq \min_{P_k} \max_{\lambda \in \sigma(A)} |P_k(\lambda)| \, \|\mathbf{e}^{(0)}\|_A \,.$$

Rate of convergence of CGM

Proof. Cont. Now, assume that, for the spectrum $\sigma(A)$, we know the largest and the smallest eigenvalues, or some lower and upper bounds, say, $0 < m \le \lambda \le M$. Then the following minimization problem, on the class of polynomials of degree k, arises:

$$P_k(0) = 1, \qquad \max_{x \in [m,M]} |P_k(x)| \to \min .$$

This problem has a classical solution $P_k^* = T_k^*$, where T_k^* is the Chebyshev polynomial on the interval $[m, M]$, which is obtained by dilation and translation of the standard Chebyshev polynomial T_k given on the interval $[-1, 1]$:

$$T_k(x) = \cos k\theta, \qquad x = \cos\theta, \qquad \theta \in [0, \pi] .$$

One can show that $|T_k^*(x)| \le 2\rho^k$ on the interval $[m, M]$, hence the rate of convergence of CGM admits the following estimate:

$$\|\mathbf{e}^{(k)}\|_A \le 2\rho^k \|\mathbf{e}^{(0)}\|_A, \qquad \rho = \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}} < 1, \qquad \sigma(A) \in [m, M] .$$

□