Optimality of CG: \((\text{CG minimizes } \|e\|_A\text{ at each step})\)

**Thm 38.2 (Trefethen & Bau)** Let the CG iteration be applied to \(Ax = b\), \(A\) symmetric positive definite. If the iteration has not already converged (i.e., if \(r_{m-1} \neq 0\)) then \(x_m\) is the unique point in \(K_m\) that minimizes \(\|e_m\|_A\). The convergence is monotonic, \(\|e_m\|_A \leq \|e_{m-1}\|_A\)

and \(e_m = 0\) is achieved for some \(m \leq m\) (in exact arithmetic.)

**pf:** \(\text{Thm 38.1 } \Rightarrow x_m \in K_m\)

Let \(x = x_m - \Delta x \in K_m\). Note \(\Delta x = x_m - x \in K_m\)

We define \(e = x_k - x = \bar{r}_k - x_m + \Delta x = e_m + \Delta x\)

Then \(\|e\|_A^2 = (e_m + \Delta x)^T A (e_m + \Delta x)\)

\[= e_m^T A e_m + \Delta x^T A \Delta x + 2 e_m^T A \Delta x\]

\[= e_m^T A e_m + \Delta x^T A \Delta x + 2 \bar{r}_m^T \Delta x\]

\[= e_m^T A e_m + \Delta x^T A \Delta x\]

\[\geq e_m^T A e_m \quad \text{since } \Delta x \in K_m\]

with equality iff \(\Delta x = 0\) i.e. \(x = x_m\)

\(K_m \subseteq K_{m+1} \Rightarrow \|e_{m+1}\|_A \leq \|e_m\|_A + K_m = \|r_m\| \Rightarrow e_m = 0\) for some \(m \leq m\)
CG and polynomial approximation

Note Thm 38.1 \( \Rightarrow \) \( x_m = q_m(A) b \in \mathcal{K}_m = \{ b, A b, \ldots, A^{m-1} b \} \)

\( m \)-1st degree poly. in \( A \)

So \( e_m = x_* - x_m = x_* - q_m(A) b \) \[ (x_* = A^{-1} b) \]

\[ (x_0 = 0) \]

\[ = x_* - x_0 - q_m(A) A^*(x_* - x_0) \]

\[ = e_0 - A q_m(A) e_0 \] (since \( A^k A = A A^k \))

\[ = p_m(A) e_0 \]

Where \( p_m(x) := 1 - x q_m(x) \) is a polynomial of degree \( m \) with \( p_m(0) = 1 \)

i.e. \( p_m(x) = 1 + c_1 x + c_2 x^2 + \ldots + c_m x^m \)

Def: \[ \mathcal{P}_m = \{ p_m(x) | p_m(0) = 1, p_m(x) = \text{poly of degree } m \} \]

**CG Approximation Problem**: Find \( p_m \in \mathcal{P}_m \)

s.t. \( \| p_m(A) e_0 \| = \text{minimum} \)
Thm 38.3 (Trefethen & Bau) If the CG iteration has not already converged before step $m$ (i.e., $r_{m-1} = 0$), then $\|p_m(A)e_0\| = \text{minimum}$ has a unique solution $p_m \in P_m$ and the iterate $x_m$ has error $e_m = p_m(A)e_0$ for this same polynomial $p_m$.

Consequently, we have

$$\frac{\|e_m\|_A}{\|e\|_A} = \inf_{p \in P_m} \frac{\|p(A)e_0\|_A}{\|e\|_A} \leq \inf_{p \in P_m} \max_{\lambda \in \Lambda(A)} |p(\lambda)|$$

where $\Lambda(A)$ denotes the spectrum (= set of all eigenvalues) of $A$.

As discussed above

$P \supseteq \text{Thm 38.1} \Rightarrow e_m = p(A)e_0$ for some $p \in P_m$

Then Thm 38.2 \Rightarrow the minimizer $e_m$ is unique

i.e., $\exists$ unique $p_m \in P_m$ s.t.

$$\|e_m\|_A = \|p_m(A)e_0\|_A = \|p(A)e_0\|_A \quad \forall p \in P_m$$

So

$$\frac{\|e_m\|_A}{\|e\|_A} = \inf_{p \in P_m} \frac{\|p(A)e_0\|_A}{\|e\|_A}$$

Next, recall $A$ sym pos def $\Rightarrow A_{jj} = \gamma_j, j = 1, \ldots, m$

where $\gamma_j \geq 0$ and $\{\gamma_j\}$ form an orthonormal basis for $\mathbb{R}^m$, i.e., $\langle v_i, \ldots, v_m \rangle = \sum_{i,j} v_i^T A v_j$ and $A_{ij} = \delta_{ij}$

Then $e_0 = \sum_{j=1}^m a_j v_j$ and $p(A)e_0 = \sum_{j=1}^m a_j p(\gamma_j)v_j = \sum_{j=1}^m a_j \phi(\gamma_j) v_j$. 
\[ ||e_0||^2_A = e_0^T A e_0 = \sum_{i,j=1}^{m} a_i a_j \lambda_i \lambda_j = \sum_{i,j=1}^{m} a_i a_j \lambda_i \delta_{ij} = \sum_{j=1}^{m} a_j^2 \lambda_j \]

Likewise
\[ ||p(A)e_0||^2_A = \sum_{j=1}^{m} a_j^2 \lambda_j (p(\lambda_j))^2 \]

\[ \frac{||p(A)e_0||^2_A}{||e_0||^2_A} = \frac{\sum_{j=1}^{m} a_j^2 \lambda_j (p(\lambda_j))^2}{\sum_{j=1}^{m} a_j^2 \lambda_j} \leq \max_{\lambda \in \Lambda(A)} |p(\lambda)|^2 \]

This gives the inequality.

Rate of convergence

Note that Theorem 3.8.3 says that the rate of convergence is determined by the \( \lambda_i \)'s.

A "good" spectrum is one on which \( p_m(\lambda_j) \) are small.

This happens if the eigenvalues are "well-grouped":

1. in small clusters
2. relatively far from 0
Theorem 38.4: If A has only m distinct eigenvalues, then the CG iteration converges in at most m steps.

(The trick in these proofs is to choose a good polynomial.

Proof: Let \( p(x) = \prod_{j=1}^{m} (1 - \frac{x}{\lambda_j}) = \prod_{j=1}^{m} (1 - \frac{x}{\lambda_1})(1 - \frac{x}{\lambda_2}) \cdots (1 - \frac{x}{\lambda_m}) \)

Then \( p \in \mathbb{P}_m \) since \( p(0) = 1 \) and \( \deg p = m \).

If \( \lambda_1, \ldots, \lambda_m \) are the m distinct eigenvalues of A, then \( p(\lambda_j) = 0 \) for any \( \lambda_j \in \Lambda(A) \).

\( \min_{\varphi \in \mathbb{P}_m} \max_{\lambda \in \Lambda(A)} |\varphi(\lambda)| = \max_{\varphi \in \mathbb{P}_m} \min_{\lambda \in \Lambda(A)} |\varphi(\lambda)| = 0 \)

and so \( ||e_m||_A = 0 \) and CG has converged.
Another result of this type is given in
C. Golub + C. Van Loan Matrix Computations 3rd edition,
p. 530

Thm 10.2.5 If $A = I + B$ is an $m \times m$ symmetric positive definite matrix and $\text{rank}(B) = r$, then CG converges in at most $r+1$ steps.

Proof: $K_m = \langle b, A b, A^2 b, \ldots, A^{m-1} b \rangle$

$= \langle b, (I+B) b, (I+B)^2 b, \ldots, (I+B)^{m-1} b \rangle$

$= \langle b, b + B b, b + 2B b + B^2 b, \ldots, b + \ldots + B^{m-1} b \rangle$

$= \langle b, B b, B^2 b, \ldots, B^{m-1} b \rangle$

at most $r+1$ of these can be linearly independent

i.e. $\dim K_m \leq r+1$ for all $m \leq m$

we must have $e_m = 0$ for some $m > r+1$

Remark: For Fornberg's method we get a matrix of this form where $B$ is the discretization of a compact integral operator $Q$. Compact operators are "nearly finite rank $r"$ more later."
For compact operators the eigenvalues converge to 0. If the kernel of an integral operator is smooth enough, we can say something about the rate of convergence of the eigenvalues to 0. The matrix $B$ which is the discretization of the operator $R$ often inherits the properties of $R$. For instance, if the kernel $R$ is analytic then the eigenvalues of $B$ are roughly of the form $\mu_j \approx r^j$ for some $0 < r < 1$.

The eigenvalues of $I + B$ are of the form

$$\lambda_j = 1 + \mu_j \approx 1 + cr^j$$

and the cluster around 1 as $j \to \infty$.

Using

$$\rho(x) = \prod_{j=1}^{m} \left(1 - \frac{x}{\lambda_j}\right) = \prod_{j=1}^{m} \frac{\lambda_j - x}{\lambda_j} = \prod_{j=1}^{m} \frac{1 + \mu_j + x}{1 + \mu_j}$$

$$\max_{\lambda \in \Lambda} |\rho(x)| \leq \max_{\lambda \in \Lambda} \frac{|C|}{1 - cr^j} \prod_{j=1}^{m} \left|1 + \mu_j - 1 - \mu_k\right|$$

$$\leq \max_{\lambda \in \Lambda} \frac{|C|}{1 - cr^j} \prod_{j=1}^{m} \left|1 + \mu_j - 1 - \mu_k\right|$$

$$\approx 1 + m$$
\[ \leq \max C \prod_{j=1}^{m} |\mu_j - \mu_k| \]

\[ \leq \max C \prod_{j=1}^{m} (|\mu_j| + |\mu_k|) \]

\[ \leq \max 2C \prod_{j=1}^{m} |\mu_j| \]

\[ \leq \max 2C \prod_{j=1}^{m} r_j \]

\[ = \max 2C \prod_{j=1}^{m} \frac{m}{2} \]

\[ = \max 2C \prod_{j=1}^{m} \frac{m(m+1)}{2} \]

\[ = \max 2C \prod_{j=1}^{m} \frac{m^2}{2} \]

\[ \leq C \frac{m^2}{2} \]

\[ \hat{r} = \sqrt{r} < 1 \]

This can be very rapid convergence.

E.g. if \( \hat{r} \approx \frac{1}{2} \) \( C = 1 \)

for \( m = 11 \)

\[ \frac{1}{l_1} \approx 2^{-7} = 2^{-4.9} = (2^{-10})^{4.9} \approx 10^{-15} \]

\[ 2^{10} = 1024 \]
Next we need

Chebyshev polynomials, \( T_n(x) \) of degree \( n \geq 0 \).

\[
T_n(x) := \cos(n \cos^{-1} x), \quad -1 \leq x \leq 1, \quad n = 0, 1, 2, \ldots
\]

Let \( \theta = \cos^{-1} x \), i.e. \( x = \cos \theta \) \( 0 \leq \theta \leq \pi \).

Then

\[
T_0(x) = \cos(0 \cdot \theta) = 1
\]

\[
T_1(x) = \cos(\theta) = x
\]

\[
T_2(x) = \cos(2\theta) = 2\cos^2 \theta - 1 = 2x^2 - 1
\]

\[
T_3(x) = \cos(3\theta) = 4\cos^3 \theta - 3\cos \theta
\]

\[
T_4(x) = \cos(4\theta) = 8\cos^4 \theta - 8\cos^2 \theta + 1
\]

\[
T_n(x) = \cos(n\theta) = \cos(n \cos^{-1} x)
\]

\[
T_{n+1}(x) = \cos((n+1)\theta) = \cos(n \theta + \theta)
\]

\[
= \cos(n \theta) \cos \theta + \sin(n \theta) \sin \theta
\]

\[
= T_n(x) \cdot x + \sin(n \theta) \sin \theta
\]

\[
T_n(x) = 2x T_{n-1}(x) - T_{n-2}(x)
\]

or

\[
T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x)
\]

This is a three-term recurrence formula,
Chebyshev polynomials are examples of orthogonal polynomials which arise frequently in classical analysis and areas of numerical analysis such as numerical integration (Gaussian quadrature), approximation theory, numerical odes + pdes, ... Polynomial approximation is often useful in obtaining e.g. estimates of order of error in finite element methods, ..., and as we will see, estimates of rates of convergence for iterative methods for solving $Ax = b$.

Orthogonal polynomials on $[-1, 1]$ are polynomials $p_m(x)$ of degree $m \geq 0$ which satisfy orthogonality relations

$$\int_{-1}^{1} p_m(x) p_n(x) \omega(x) dx = 0, \quad m \neq n$$

where $\omega(x)$ is a given weight function satisfying $\omega(x) \geq 0 \; x \in [-1, 1]$ and $0 < \int_{-1}^{1} \omega(x) dx < \infty$.

Problem: Show that the Chebyshev polynomials $T_m(x)$ satisfy orthogonality relations with $\omega(x) = \frac{1}{\sqrt{1-x^2}}$.

See e.g. A. Iserles A First Course in Numerical Analysis of Differential Equations sec 3.1, L. Trefethen + D. Bau Numerical Linear Algebra Lecture 37, ...
Orthogonal polynomials, such as Chebyshev, Legendre, Laguerre, Hermite, ... polynomials satisfy recurrence formulas like the one above; see any Mathematical Handbook for a compilation of these and other identities for these special functions. The polynomials can be generated by these recurrence formulas, e.g., for Chebyshev polynomials

\[ T_0(x) = 1 \]
\[ T_1(x) = x \]

and

\[ T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x) \]

gives

\[ T_2(x) = 2xT_1(x) - T_0(x) = 2x^2 - 1 \]
\[ T_3(x) = 2x(T_2(x)) - T_1(x) = 4x^3 - 3x \]

\[ \vdots \]

Problem: Find \( T_4(x) \) and plot it. Make a MATLAB plot.
The following change of variables is useful:

\[ x = \cos \theta = \frac{1}{2} (e^{i \theta} + e^{-i \theta}) = \frac{1}{2} (z + z^{-1}) \quad \therefore \quad z = e^{i \theta} \]

Note: \( f(z) = \frac{1}{2} (z + z^{-1}) \) is the Joukowskki transform.

\[ \begin{array}{c}
\frac{1}{2}(p+\frac{1}{p}) \\quad f \quad \frac{1}{2}(p+\frac{1}{p}) \\
\end{array} \]

\[ f(1) = e^{i \theta} \quad p > 1 \text{ fixed}, \quad 0 \leq \theta \leq 2\pi \]

\[ f \text{ maps the slit } [-1,1] \] covered twice.

The ellipse \( f(p e^{i \theta}) = \frac{1}{2} (p e^{i \theta} + \frac{1}{p} e^{-i \theta}) \)

\[ = \frac{1}{2} (p + \frac{1}{p}) \cos \theta + \frac{i}{2} (p - \frac{1}{p}) \sin \theta \]

with major axis \( [-\frac{1}{2}(p+\frac{1}{p}), \frac{1}{2}(p+\frac{1}{p})] \)

and minor axis \( [\frac{1}{2}(p-\frac{1}{p}), \frac{1}{2}(p-\frac{1}{p})] \)

The radial lines \( 1 < p < \infty \), \( \theta \) fixed
are mapped to hyperbolas.

\( f \) maps \( |z| > 1 \) and \( |z| < 1 \) each \( -1 \) onto \( \mathbb{C} \setminus [-1,1] \), with simple poles at \( z = 0, \infty \).

Remark: The Joukowskii maps is a good test case for exterior conformal mapping methods.
Continuing with our change of variables

\[ x = \frac{1}{2} (z + z^{-1}) = f(z) \]

gives

\[ z^2 - 2xz + 1 = 0 \]

and

\[ z = \frac{2x \pm \sqrt{4x^2 - 4}}{2} = x \pm \sqrt{x^2 - 1} \]

Note

\[ x - \sqrt{x^2 - 1} = \frac{x^2 - x^2 + 1}{x + \sqrt{x^2 - 1}} = \frac{1}{x + \sqrt{x^2 - 1}} \]

We take \( \bar{z} = x + \sqrt{x^2 - 1} \) then \( z^{-1} = \frac{x}{x - \sqrt{x^2 - 1}} \)

and \( f(i) = \frac{1}{2} (i - i) = 0 \), \( f(-i) = \frac{1}{2} (-i + i) = 0 \)

If \( y > 1 \), then \( y = f(z) = \frac{1}{2} (z + z^{-1}) \)

Then

\[ z = y + \sqrt{y^2 - 1} > 1 \]

Note also \( T_m(x) = \cos(m \cdot \theta) \)

\[ z = e^{i \theta} \]

\[ \theta = -i \ln z \]

\[ = \cos(-im \ln z) \]

\[ = \frac{1}{2} (e^{im \ln z} + e^{-im \ln z}) = \cosh (m \ln z) \]

\[ = \frac{1}{2} (z^m + z^{-m}) \]

\[ = \frac{1}{2} (\left( x + \sqrt{x^2 - 1} \right)^m + (x - \sqrt{x^2 - 1})^m) \]

\[ = \text{poly of degree } m \text{ in } x \text{ (not so obvious)} \]

(So \( T_m(x) \) is analytic for all complex \( x \).)
For $A \in \mathbb{R}^{m \times m}$ positive definite $k = k_2(A) = \| A \|_2 \| A^{-1} \|_2 = \frac{\lambda_{\max}}{\lambda_{\min}} \geq 1$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $A$:

$0 < \lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m = \lambda_{\max}$

i.e., $k \lambda_1 = \lambda_m$ and so the distance of the eigenvalues of $A$ vary at most by a factor of $k$ from the origin.

**Theorem 38.5 (Trefethen & Bau)** Let the CG iteration be applied to a symmetric positive definite matrix problem $Ax = b$, where $A$ has 2-norm condition number $k := k_2(A)$. Then the $A$-norms of the errors satisfy

$$\frac{\|e_m\|_A}{\|e_0\|_A} \leq \frac{2}{\left(\frac{\sqrt{k} - 1}{\sqrt{k} + 1}\right)^m + \left(\frac{\sqrt{k} - 1}{\sqrt{k} + 1}\right)^{-m}} \leq 2 \left(\frac{\sqrt{k} - 1}{\sqrt{k} + 1}\right)^m$$

Note to guarantee $\frac{\|e_m\|_A}{\|e_0\|_A} \leq \epsilon = "tolerance" \ll 1$

we need to take at least $m$ steps such that

$$\ln k + m \ln \left(\frac{\sqrt{k} - 1}{\sqrt{k} + 1}\right) < \ln \epsilon$$

since $\ln \left(\frac{\sqrt{k} - 1}{\sqrt{k} + 1}\right) = \ln \left(\frac{1 - \sqrt{k}}{1 + \sqrt{k}}\right) \sim -\frac{3}{\sqrt{k}}$, $k \rightarrow \infty$

we need $m \geq \frac{\sqrt{k}}{2} \ln \epsilon + \frac{\sqrt{k}}{2} \ln 2 = O(\sqrt{k})$ steps.
Proof: By Thm 38.3 it is enough to find a poly.

\[ p_m \in P_m \text{ s.t. } \max_{x \in [\lambda_{\min}, \lambda_{\max}]} |p(x)| = \frac{2}{\left(\lambda_{\max} - \lambda_{\min}\right)^m} \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)^m \]

Let \[ p_m(x) = \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} + \frac{2x}{\lambda_{\max} - \lambda_{\min}} \]

where \[ \gamma = \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} = \frac{x + 1}{x - 1} > 1 \]

This is the scaled and shifted Chebyshev poly.

Note \[ p_m(0) = \frac{T_m(0)}{T_m(\gamma)} = 1 \]

and \[ p_m(x) \text{ is of degree } m \]

\[ p_m \in P_m \]

\[ \Rightarrow \text{ we want } 0 < \varepsilon \leq \frac{2}{T_m(\lambda_{\max})} \]

\[ \text{to make this small} \]

\[ \text{i.e. A good \quad \text{ "approx. of 0"} \quad \text{on } [\lambda_{\min}, \lambda_{\max}] \}

Note also

\[ -1 \leq x - \frac{2x}{\lambda_{\max} - \lambda_{\min}} = \frac{\lambda_{\max} + \lambda_{\min} - 2x}{\lambda_{\max} - \lambda_{\min}} \leq 1 \]

\[ (x = \lambda_{\max}) \]

for \[ x \in [\lambda_{\min}, \lambda_{\max}] \]
Since \(|T_n(x)| = |\cos(n \cos^{-1} \theta)| \leq 1\) for \(x \in [-1, 1]\), we have \(|P_n(x)| \leq \frac{1}{|T_n(x)|}\) for \(x \in [\sin(R_{\min}), \sin(R_{\max})]\).

From the properties of \(T_n\) above we have

\[
T_n(x) = \frac{1}{2} \left( 2^n x + 2^{-n} \right)
\]

where

\[
z = \frac{\frac{k+1}{k-1} + \sqrt{\left(\frac{k+1}{k-1}\right)^2 - 1}}{k-1}
\]

\[
= \frac{k+1 + \sqrt{4k}}{k-1} = \frac{k+1 + 2\sqrt{k}}{k-1}
\]

\[
= \frac{k+1 + 2\sqrt{k} + 1}{k-1} = \frac{(\sqrt{k}+1)^2}{(\sqrt{k}+1)(\sqrt{k}-1)} = \frac{\sqrt{k}+1}{\sqrt{k}-1}
\]

\[
T_n(x) = \frac{1}{2} \left( \left(\frac{\sqrt{k}+1}{\sqrt{k}-1}\right)^n + \left(\frac{\sqrt{k}+1}{\sqrt{k}-1}\right)^{-n} \right)
\]

and we are finished. \(\square\)