Optimality of CG: (CG minimizes $\|e\|_A$ 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 $e_n = 0$ is achieved for some $n \leq m$ (in exact arithmetic.).

---

pf: Thm 38.1 $\Rightarrow$ $x_m \in K_m$

Let $x = x_m - \Delta x \in K_m$. Note $\Delta x = x_m - x \in K_m$ arbitrary

and error $e = x_* - x = x_* - x_m + \Delta x = e_m + \Delta x$

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

$$= e_m^T A e_m + \Delta x^T A \Delta x + 2 e_m^T A \Delta x$$

$r_m = Ax_* - Ax_m = b - Ax_m$

$$= e_m^T A e_m + \Delta x^T A \Delta x + 2 \underbrace{r_m^T \Delta x}_{=0}$$

0 by Thm 38.1 since $\Delta x \in K_m$

$$= e_m^T A e_m + \Delta x^T A \Delta x$$

$$\geq e_m^T A e_m \quad \text{since } A \text{ pos. def.}$$

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

$K_m \subseteq K_{m+1} \Rightarrow \|e_{m+1}\|_A < \|e_m\|_A$ + $K_m = \mathbb{R}^m \Rightarrow e_n = 0$ for some $n \leq m$ (qed)

CG and polynomial approximation

Note Thm 38.1 $\Rightarrow$ $x_m = \underbrace{q_m}_{\substack{\uparrow \\ m-1\text{st degree poly. in } A}}(A)b \in K_m = \langle b, Ab, \dots, A^{m-1}b \rangle$

$$\text{So} \quad e_m = x_* - x_m = x_* - q_m(A)b \qquad (x_* = A^{-1}b)$$

$$= x_* - x_0 - q_m(A)A(x_* - x_0) \qquad (x_0 = 0)$$

$$= e_0 - A\,q_m(A)\,e_0 \qquad (\text{since } A^kA = AA^k)$$

$$= p_m(A)\,e_0$$

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

i.e. $p_m(x) = 1 + c_1 x + c_2 x^2 + \dots + c_m x^n$

Def: $\mathbb{P}_m = \{ p_m(x) \mid p_m(0) = 1,\ p_m(x) = \text{poly of degree } m \}$

$$\boxed{\begin{array}{l} \text{CG Approximation Problem: Find } p_m \in \mathbb{P}_m \\ \text{s.t.} \quad \| p_m(A)\,e_0 \| = \text{minimum} \end{array}}$$

Thm 38.3 (Trefethen & Bau) If the CG iteration has not already converged before step m (i.e. $r_{m-1} \neq 0$), then $\|p_m(A)e_0\| = $ 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

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

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

Pf: As discussed above 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

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

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

Next, recall A sym pos def $\Rightarrow$ $Av_j = \lambda_j v_j \quad j = 1, \dots, m$ where $\lambda_j \geq 0$ and v_j form an orthonormal basis for $\mathbb{R}^m$, i.e. $\langle v_1, \dots, v_m \rangle = \mathbb{R}^m$ and $v_i^T v_j = \delta_{ij}$

Then $e_0 = \sum_{j=1}^m a_j v_j$ and $p(A)e_0 = \sum_{j=1}^m a_j p(A)v_j = \sum_{j=1}^m a_j p(\lambda_j) v_j$.

$$\therefore \|e_0\|_A^2 = e_0^T A e_0 = \sum_{i,j=1}^{m} a_i a_j v_i A v_j = \sum_{i,j=1}^{m} a_i a_j \lambda_j v_i^T v_j$$

$$= \sum_{i,j=1}^{m} a_i a_j \lambda_j \delta_{ij} = \sum_{j=1}^{m} a_j^2 \lambda_j$$

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

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

This gives the inequality. QED

Rate of convergence

Note Thm 38.3 says that the rate of convergence is determined by the λ_j's.

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

This happens if the e'values are "well-grouped":

1.) in small clusters

or 2.) relatively far from 0

1.) e'values "perfectly clusted" in m groups

<u>Thm 38.4</u> If A has only m distinct e'values, Then the CG iteration converges in at most m steps

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

<u>Pf:</u> Let $p(x) = \prod_{j=1}^{m} \left(1 - \frac{x}{\lambda_j}\right) = \left(1-\frac{x}{\lambda_1}\right)\left(1-\frac{x}{\lambda_2}\right)\cdots\left(1-\frac{x}{\lambda_m}\right)$

Then $p \in P_m$ since $p(0)=1$ and $\deg p = m$.

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

$$\therefore \inf_{q \in P_m} \max_{\lambda \in \Lambda(A)} |q(\lambda)| \leq \max_{\lambda \in \Lambda(A)} |p(\lambda)| = 0$$

and so $\|e_m\|_A = 0$ and CG has converged.

qed.

Another result of this type is given in G. Golub + C. VanLoan Matrix Computations 3rd edition, p. 530

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

---

Pf:

$$
\begin{aligned}
K_n &= \langle b, Ab, A^2 b, \ldots, A^{n-1} b \rangle \\
&= \langle b, (I+B)b, (I+B)^2 b, \ldots, (I+B)^{n-1} b \rangle \\
&= \langle b, b + Bb, b + 2Bb + B^2 b, \ldots, b + \ldots + B^{n-1} b \rangle \\
&= \langle \underbrace{b, Bb, B^2 b, \ldots, B^{n-1} b}_{\substack{\text{at most } r+1 \text{ of these} \\ \text{can be linearly independent}}} \rangle
\end{aligned}
$$

i.e. $\dim K_n \le r+1 \quad \forall\, n \le m$

$\therefore$ we must have $e_n = 0$ for some $n \le r+1$.

qed

Remark For Fornberg's method we get a matrix of this form where B is the discretization of a compact integral operator R. 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 e'values 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 μ_j s.t

$$|\mu_j| \le r^j \quad \text{for some } 0<r<1.$$

$\therefore$ the e'values of $I+B$ are of the form

$$\lambda_j = 1 + \mu_j \approx 1 \pm c r^j$$

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

Using

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

$$\therefore \max_{\lambda \in \Lambda(A)} |p_m(\lambda)| \le \max_{\lambda \in \Lambda(A)} \frac{C}{1-cr} \prod_{j=1}^{m} |1+\mu_j - \lambda|$$

$$\le \max_{\substack{\lambda = 1+\mu_k \\ k>m}} \frac{C}{1-cr} \prod_{j=1}^{m} |1+\mu_j - 1 - \mu_k|$$

$$\le \max_{k>m} C \prod_{j=1}^{m} |\mu_j - \mu_k|$$

$$\le \max_{k>m} C \prod_{j=1}^{m} \left(|\mu_j| + |\mu_k| \right)$$

$$\le \max_{k>m} 2C \prod_{j=1}^{m} |\mu_j|$$

$$\le \max_{k>m} 2C \prod_{j=1}^{m} r^j$$

$$= \max 2C\, r^{\sum_{j=1}^{m} j}$$

$$= \max 2C\, r^{\frac{m(m+1)}{2}}$$

$$\le C\, \hat{r}^{m^2} \qquad \hat{r} = \sqrt{r} < 1$$

This can be <u>very</u> rapid convergence.

e.g. if $\hat{r} \approx \frac{1}{2}$ $\quad C \approx 1$

for $m = 7$

$$\frac{\|e_m\|_A}{\|e_0\|_A} \approx 2^{-7^2} = 2^{-49} = \left(2^{-10}\right)^{4.9} \approx 10^{-15} \approx \underline{\varepsilon_{mach}}!$$

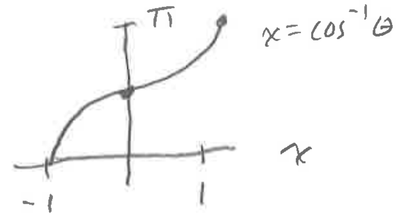
$$2^{10} = 1024$$

Next we need

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

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

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



Then $\quad T_n(x) = \cos(n\theta)$

$\underline{n=0}$ $\quad T_0(x) = \cos(0 \cdot \theta) = 1$

$\underline{n=1}$ $\quad T_1(x) = \cos(\theta) = x$

$\underline{n=2}$ $\quad T_2(x) = \cos(2\theta) = 2\cos^2\theta - 1 = 2x^2 - 1$

$\vdots$

$\underline{n \geq 1}$ $\quad$ Using $\cos(\alpha \pm \beta) = \cos\alpha \cos\beta \mp \sin\alpha \sin\beta$

$$\begin{aligned} T_{n\pm1}(x) &= \cos((n\pm1)\theta) = \cos(n\theta \pm \theta) \\ &= \cos(n\theta)\cos\theta \mp \sin(n\theta)\sin\theta \\ &= T_n(x) \cdot x \mp \sin(n\theta)\sin\theta \end{aligned}$$

$$\therefore \quad T_{n+1}(x) + T_{n-1}(x) = 2x\,T_n(x)$$

or

$$\boxed{T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x)}$$

This is a <u>three-term recurrence formula</u>, 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 $p_n(x)$ of degree $n \geq 0$ which satisfy orthogonality relations

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

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

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

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

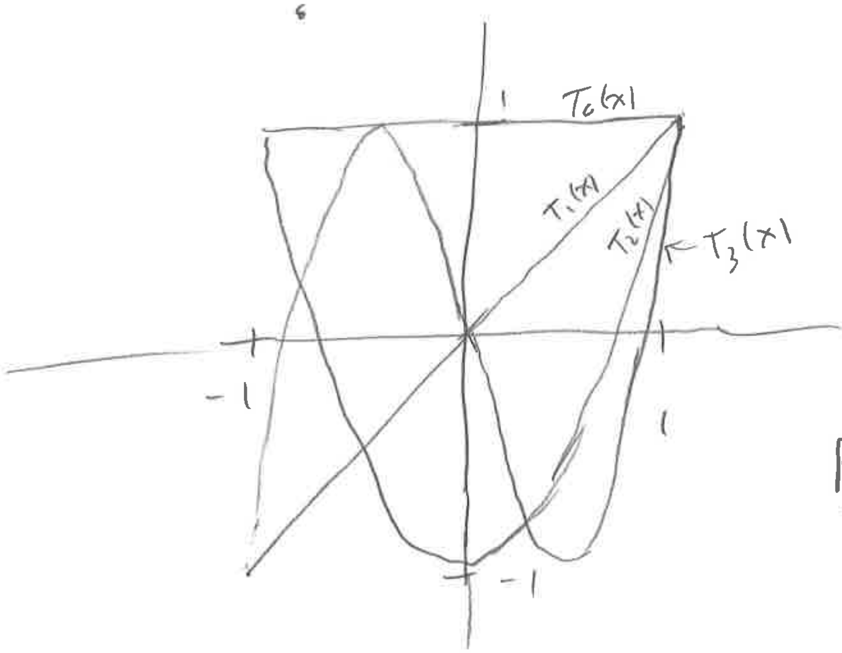
$$T_0(x) = 1$$
$$T_1(x) = x$$

and

$$T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x)$$

gives

$$T_2(x) = 2x\,T_1(x) - T_0(x) = 2x^2 - 1$$
$$T_3(x) = 2x(T_2(x)) - T_1(x) = 4x^3 - 3x$$
$$\vdots$$

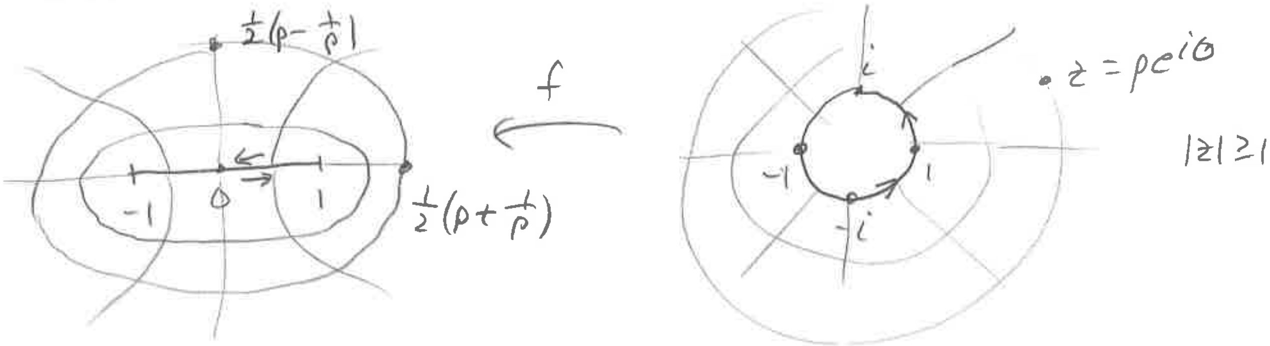


<u>Problem</u>: Find $T_4(x)$ and plot it. Make a MATLAB plot.

The following change-of-variables is useful:

$$x = \cos\theta = \tfrac{1}{2}(e^{i\theta} + e^{-i\theta}) = \tfrac{1}{2}(z + z^{-1}), \quad z = e^{i\theta}$$

Note: $f(z) := \frac{1}{2}(z + z^{-1})$ is the Joukowski transformation



f maps the unit circle $|z|=1$, $z = e^{i\theta}$ to the slit $[-1,1]$ covered twice.

f maps the circles $z = \rho e^{i\theta}$ $\quad \rho > 1$ fixed, $0 \le \theta \le 2\pi$

the ellipse $f(\rho e^{i\theta}) = \frac{1}{2}(\rho e^{i\theta} + \frac{1}{\rho} e^{-i\theta})$

$$= \tfrac{1}{2}(\rho + \tfrac{1}{\rho}) \cos\theta + \tfrac{i}{2}(\rho - \tfrac{1}{\rho}) \sin\theta$$

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

and minor axis $i[-\frac{1}{2}(\rho - \frac{1}{\rho}), \frac{1}{2}(\rho - \frac{1}{\rho})]$.

The radial lines $1 < \rho < \infty$, θ fixed are mapped to hyperbolas

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

Remark: The Joukowski maps is a good test case for exterior conformal mapping methods.

Continuing with our change-of-variables

$$x = \tfrac{1}{2}(z + z^{-1}) =: f(z)$$

gives $\quad z^2 - 2xz + 1 = 0$

and $\quad z = \dfrac{2x \pm \sqrt{4x^2 - 4}}{2} = x \pm \sqrt{x^2 - 1}$

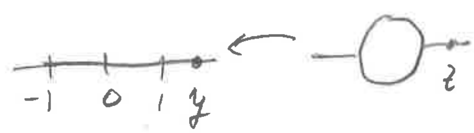
Note $\quad x - \sqrt{x^2-1} = \dfrac{x^2 - x^2 + 1}{x + \sqrt{x^2-1}} = \dfrac{1}{x + \sqrt{x^2-1}}$.

We take $z = x + \sqrt{x^2-1}$ then $z^{-1} = x - \sqrt{x^2-1}$

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

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

then $\boxed{z = y + \sqrt{y^2 - 1} > 1}$



Note also $T_m(x) = \cos(m\theta)$ $\qquad z = e^{i\theta}$, $\theta = -i \ln z$

$= \cos(-im \ln z)$

$= \tfrac{1}{2}\left(e^{m \ln z} + e^{-m \ln z}\right) = \cosh(\ln z)$

$= \tfrac{1}{2}\left(z^m + z^{-m}\right)$

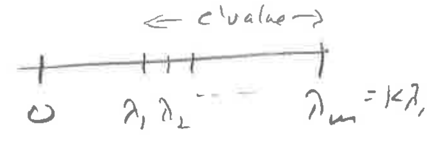
$= \tfrac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^m + \left(x - \sqrt{x^2-1}\right)^m\right)$

(= poly of degree m in x ← not so obvious)

(So $T_m(x)$ is analytic for all complex x.)

For $A \in \mathbb{R}^{m \times m}$ positive definite $\kappa = \kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \dfrac{\lambda_{max}}{\lambda_{min}} \geq 1$
where λ_{max} and λ_{min} are the largest and smallest e'values of A: $0 < \lambda_{min} = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m = \lambda_{max}$.

i.e. $\kappa \lambda_1 = \lambda_m$ and so the distance of the e'values of A vary at most by a factor of κ from the origin:



<u>Thm 38.5</u> (Trefethen & Bau) Let the CG iteration be applied to a symmetric positive definite matrix problem $Ax = b$, where A has 2-norm condition number $\kappa := \kappa_2(A)$. Then the A-norms of the errors satisfy

$$\frac{\|e_n\|_A}{\|e_0\|_A} \leq \frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{n} + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-n}} \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{n}$$

<u>Note</u> to guarantee $\dfrac{\|e_n\|_A}{\|e_0\|_A} \leq \varepsilon$ = "tolerance" $\ll 1$

we need to take at least n steps s.t.

$$\ln 2 + n \ln\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) < \ln \varepsilon .$$

since $\ln\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) = \ln\left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\right) \sim -\frac{2}{\sqrt{\kappa}}$, $\kappa \to \infty$

we need $n \gtrsim -\frac{\sqrt{\kappa}}{2} \ln \varepsilon + \frac{\sqrt{\kappa}}{2} \ln 2 = \underline{O(\sqrt{\kappa}) \text{ steps}}$

(Recall for steepest descent the conv. factor was $\frac{\kappa-1}{\kappa+1}$ requiring $O(\kappa)$ steps)

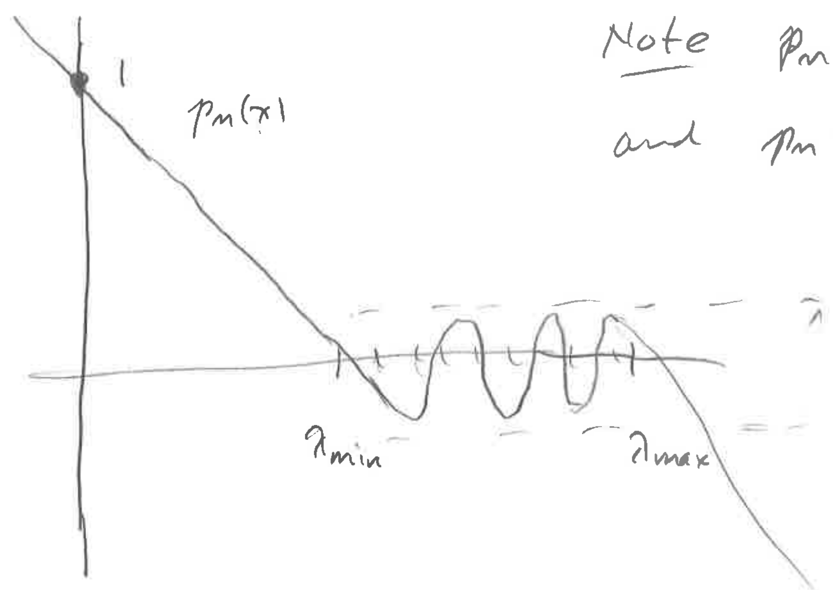
Pf By Thm 38.3 it is enough to find a poly.

$p_m \in P_m$ s.t. $\max_{\lambda \in [\lambda_{min}, \lambda_{max}]} |p(\lambda)| = \dfrac{2}{\left(\frac{\sqrt{k}+1}{\sqrt{k}-1}\right)^m + \left(\frac{\sqrt{k}+1}{\sqrt{k}-1}\right)^{-m}}$

Let $p_m(x) = T_m\left(\gamma - \dfrac{2x}{\lambda_{max} - \lambda_{min}}\right) / T_m(\gamma)$

where $\gamma := \dfrac{\lambda_{max} + \lambda_{min}}{\lambda_{max} - \lambda_{min}} = \dfrac{k+1}{k-1} > 1.$

This is the scaled and shifted Chebyshev poly



Note $p_m(0) = \dfrac{T_m(\gamma)}{T_m(\gamma)} = 1$

and $p_m(x)$ is of degree m

$\therefore \; p_m \in P_m$

We want to make this small $\Leftarrow$ in fact $\leq \dfrac{2}{|T_m(\gamma)|}$

i.e. a good "approx. of 0" on $[\lambda_{min}, \lambda_{max}]$

Note also

$$\underset{(x=\lambda_{max})}{-1} \leq \gamma - \frac{2x}{\lambda_{max} - \lambda_{min}} = \frac{\lambda_{max} + \lambda_{min} - 2x}{\lambda_{max} - \lambda_{min}} \leq \underset{(x=\lambda_{min})}{1}$$

for $x \in [\lambda_{min}, \lambda_{max}]$

Since $|T_m(x)| = |\cos(m\cos^{-1}\theta)| \leq 1$ for $x \in [-1,1]$

we have $|p_m(x)| \leq \dfrac{1}{|T_m(\gamma)|}$ for $x \in [\lambda_{min}, \lambda_{max}]$

From the properties of T_m above we have

$$T_m(\gamma) = \frac{1}{2}(z^m + z^{-m})$$

where

$$z = \gamma + \sqrt{\gamma^2 - 1} = \frac{\kappa+1}{\kappa-1} + \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1}$$

$$= \frac{\kappa + 1 + \sqrt{(\kappa+1)^2 - (\kappa-1)^2}}{\kappa - 1}$$

$$= \frac{\kappa+1 + \sqrt{4\kappa}}{\kappa-1}$$

$$= \frac{\kappa + 2\sqrt{\kappa} + 1}{\kappa - 1} = \frac{(\sqrt{\kappa}+1)^2}{(\sqrt{\kappa}+1)(\sqrt{\kappa}-1)} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$$

$$\therefore \quad T_m(\gamma) = \frac{1}{2}\left(\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^m + \left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-m}\right)$$

and we are finished

Qed