Things to Add:

- Addition of preconditioner, M.
- Chebyshev polynomial $\longleftrightarrow \cos(n\theta)$
- Chebyshev semi-iterative method

Convergence of CG

The CG convergence analysis proceeds from the following observations.

- The kth iterate, $\underline{x}_k$ is the best possible approximation in the Krylov subspace,

$$K_k(A; \underline{b}) = span\{\underline{b}\ A\underline{b} \ldots\ A^{k-1}\underline{b}\}$$

- $\underline{x}_k$ can be expressed as a polynomial in A as

$$\underline{x}_k \;=\; c_1\underline{b} + c_2A\underline{b} + \cdots + c_kA^{k-1}\underline{b} = P_{CG}^{k-1}(A)\underline{b}, \qquad (1)$$

  where the unknown coefficients c_j are optimally determined by the conjugate gradient algorithm.

- Note that P_{CG}^{k-1} is generally an unknown polynomial but it has the special property that

$$\|\underline{x} - P_{CG}^{k-1}(A)\underline{b}\|_A \;\leq\; \|\underline{x} - P^{k-1}(A)\underline{b}\|_A \;\; \forall\, P^{k-1}(A) \;\in\; \mathbb{P}_{k-1}(A). \qquad (2)$$

- To analyze the convergence behavior, notice that the error satisfies,

$$\underline{e}_k \;=\; \underline{x} - \underline{x}_k = A^{-1}(\underline{b} - A\underline{x}_k) = A^{-1}\underline{r}_k. \tag{3}$$

- Thus the A-norm of the error, minimized by CG, is given by:

$$\begin{aligned} \|\underline{e}_k\|_A^2 \;&=\; \underline{e}_k^T A \underline{e}_k \\ &=\; \underline{r}_k^T A^{-1} \underline{r}_k \;=\; \|\underline{r}_k\|_{A^{-1}}^2 \quad . \end{aligned} \tag{4}$$

- Inserting the polynomial representation for $\underline{x}_k$ into the expression for $\underline{r}_k$, we have:

$$\begin{aligned} \underline{r}_k \;&=\; \underline{b} \,-\, A\underline{x}_k \\ &=\; \underline{b} \,-\, c_1 A\underline{b} \,-\, c_2 A^2 \underline{b} \,-\, \ldots \,-\, c_k A^k \underline{b} \quad . \end{aligned} \tag{5}$$

- Note that the degrees of freedom in (5) are represented by the c_j's.

- Thus, out of all possible polynomials having the form

$$P_1^k(t) \;=\; 1 \,+\, \gamma_1 t \,+\, \ldots \,+\, \gamma_k t^k \tag{6}$$

  (i.e., those satisfying $P_1^k(0) = 1$), the conjugate gradient algorithm constructs the one which minimizes $\|\underline{e}_k\|^2$,

$$\begin{aligned} \|\underline{e}_k\|_A^2 \;&=\; \underline{r}_k^T A^{-1} \underline{r}_k \\ &=\; \underline{b}^T (I - AP_{CG}^{k-1})^T A^{-1} (I - AP_{CG}^{k-1}) \underline{b} \\ &\le\; \underline{b}^T [P_1^k(A)]^T A^{-1} P_1^k(A) \underline{b} \quad , \end{aligned} \tag{7}$$

- To establish an upper bound on the error, we can choose the particular polynomial $P_1^k(t) = \tilde{T}_k(t)$, the *Chebyshev polynomial* of degree k which is scaled and translated to satisfy $\tilde{T}_k(0) = 1$.

- This choice is motivated by the fact that, for a given scaling (in this case that $P_1(0) = 1$), one can construct a Chebyshev polynomial which minimizes the maximum amplitude over all polynomials in $\mathbb{P}_k^1(x)$ for x in a given interval.

- In particular, consider the interval $x \in [-1, 1]$.

- On this interval,

$$p_k(x) \;:=\; \cos\Big(\, k\, \cos^{-1}(x) \Big) \;=\; \cos\,(\, k\,\theta\,) \tag{8}$$

  is a polynomial of degree k in x that clearly has extrema ± 1.

- Shifting the roots of p_k (i.e., changing the polynomial) will cause some extrema to lower and others to rise.

- The standard Chebyshev polynomial of degree k is the one that minimizes the maximum on the interval $x \in [-1, 1]$ for all polynomials of the form

$$p_k(x) \;=\; x^k + a_{k-1}x^{k-1} + \cdots + a_1 x + a_0. \tag{9}$$

  chebplot_demo.m

```
hdr
N=8;
z= [0:1000]'/1000;
theta = pi*z;
x=cos(theta);
y=sin(theta);
z=cos(N*theta);
plot3(x,0*x,0*x,'k-',lw,4,x,y,z,'r-',lw,4);
```

- Here we will consider the interval $[\lambda_1, \lambda_n]$, where $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ are the n positive eigenvalues of A.

- Our scaling requirement is that a_0=1, which implies $p_k(0) = 1$,

$$P_k^1(x) \;=\; a_k x^k \,+\, a_{k-1} x^{k-1} \,+\, \cdots \,+\, a_1 x \,+\, 1. \tag{10}$$

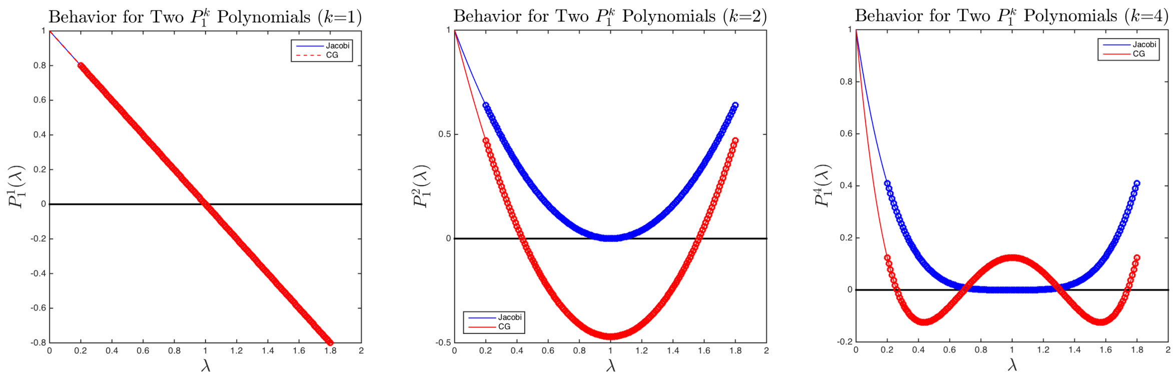

Figure 1: Comparison of error distribution for $\lambda_j \in [0.2{:}1.8]$ for error polynomials based on Jacobi iteration vs. Chebyshev distribution. The CG error distribution will be smaller than the Chebyshev one. (Why?)

- Figure 1 shows an example of error polynomials of the form $P_k^1(\lambda)$ for $\lambda \in [0{:}2]$ in which the translated/scaled Chebyshev polynomial of degree k minimizes the maximum amplitude on the interval $[\lambda_1{:}\lambda_n]$=[0.2:1.8].

- Notice that, on $[\lambda_1 : \lambda_n]$ the maximum of $|P_k^1|$ for the Chebyshev polynomial (in red, labeled "CG") is smaller than that associated with Jacobi iteration, which is given by $(1-\lambda)^k$.

- Since CG yields a *better approximation than any other polynomial of degree k* then the error will be $\le$ the error induced by a Chebyshev polynomial, and certainly better than the error associated with Jacobi iteration for any value of $k > 1$.

- The essence of the convergence proof is to use the computable maxima of the Chebyshev polynomials to bound the error for CG.

- We begin by considering a spectral decomposition of the initial residual:

$$\underline{b} \;=\; \sum_{i=1}^{n} \hat{b}_i \underline{z}_i \,, \tag{11}$$

where $\underline{z}_i$ is the eigenvector of A associated with eigenvalue λ_i normalized such that

$$\underline{z}_i^T \underline{z}_j \;=\; \delta_{ij} \,, \tag{12}$$

where δ_{ij} is the Kronecker delta.

- Because A is symmetric, it has n orthogonal eigenvectors spanning $\mathbb{R}^n$ and, consequently, there always exists a decomposition of the form (11).

- The (arbitrary) scaling of the eigenvectors is established by (12).

- We will use the following relationship shortly.

$$\|\underline{x}\|_A^2 \;=\; \|A^{-1}\underline{b}\|_A^2 \;=\; (A^{-1}\underline{b})^T A (A^{-1}\underline{b}) \;=\; \underline{b}^T A^{-1} \underline{b} \;=\; \sum_{i=1}^{n} \frac{\hat{b}_i^2}{\lambda_i}. \tag{13}$$

- Inserting the spectral decomposition (11) of $\underline{b}$ into the error equation (7) yields

$$\|\underline{e}_k\|_A^2 \leq \left(\sum_{i=1}^{n} P_1^k(\lambda_i) \hat{b}_i \underline{z}_i \right)^T \left(\sum_{j=1}^{n} P_1^k(\lambda_j) \frac{\hat{b}_j}{\lambda_j} \underline{z}_j \right) \tag{14}$$

$$= \left(\sum_{j=1}^{n} \sum_{i=1}^{n} P_1^k(\lambda_i) P_1^k(\lambda_j) \frac{\hat{b}_i \, \hat{b}_j}{\lambda_j} \underline{z}_i^T \, \underline{z}_j \right). \tag{15}$$

From the orthonormality of the eigenvectors (12) we have:

$$\|\underline{e}_k\|_A^2 \leq \sum_{i=1}^{n} (P_1^k(\lambda_i))^2 \frac{\hat{b}_i^2}{\lambda_i} \leq \sum_{i=1}^{n} M^2 \frac{\hat{b}_i^2}{\lambda_i} = M^2 \sum_{i=1}^{n} \frac{\hat{b}_i^2}{\lambda_i} = M^2 \, \|\underline{x}\|_A^2. \tag{16}$$

- Here, M is a constant corresponding to the maximum of $P_1^k(\lambda_i)$,

$$M := \max_i |P_1^k(\lambda_i)|, \tag{17}$$

which is the bound we seek. We have

$$\frac{\|\underline{e}_k\|}{\|\underline{x}\|} \leq M = \max_i |P_1^k(\lambda_i)| \tag{18}$$

$$\leq \max_{\lambda_1 \leq \lambda \leq \lambda_n} |P_1^k(\lambda)|. \tag{19}$$

- Since P_1^k may be *any* polynomial of degree k satisfying $P_1^k(0) = 1$ we can estimate a relatively sharp bound by finding a polynomial that minimizes the right-hand side of (19).

- That is, find

$$P_1^k(\lambda) \;=\; \operatorname*{argmin}_{p \in \mathbb{P}_k^1} \max_{\lambda \in [\lambda_1 : \lambda_n]} |p(\lambda)| \tag{20}$$

- The solution to this problem, as is often the case in minimax problems, is given by a scaled and translated Chebyshev polynomial mentioned previously.

- Before proceeding with that analysis, however, we note that (18) provides a sharper estimate than given by the bounds of the minimizing polynomial.

- Specifically, if most of the eigenvalues are clustered in a small region, then a polynomial that passes through the outlying λ_is and that is also small over the clustered region would yield a tighter estimate than the Chebyshev result presented below.

- We also note that if some of the $\hat{b}_j$'s are zero then they would nominally be excluded from the sums that are present in (14), save that round-off error generally prevents their contribution from being truly void.

- A more common scenario, however, is that A has eigenvalues with multiplicity > 1.
- Assume that A has $m < n$ unique eigenvalues, $\{\,\lambda_1 \,<\, \lambda_2 \,<\, \dots \,<\, \dots \lambda_m\,\}$.
- In this case, $\underline{b}$ has an equivalent spectral decomposition

$$\underline{b} = \sum_{i=1}^{m} \hat{b}_i \underline{z}_i\,, \tag{21}$$

where $\underline{z}_i$ is an eigenvector of A associated with eigenvalue λ_i.

- Note that any linear combination of eigenvectors associated with an eigenvalue having multiplicity greater than one is also an eigenvector.
- Krylov-subspace solvers to not have a mechanism to detect this multiplicity since every matrix-vector product will simply stretch (i.e., without rotating) the original component in the invariant subspace.
- **The net result is that KSPs converge in at most $m \leq n$ iterations, modulo round-off effects.**

Chebyshev Polynomials

- We turn now to the standard estimate to bound (19).

- This is a classic minimax problem which is invariably solved by using Chebyshev polynomials, $T_k(x)$.

- We reiterate that (18) provides a *tighter* error bound because the maximum in (18) is taken over a *discrete* set of eigenvalues and this maximum will generally be smaller than the maximum found on the continuous interval $[\lambda_1, \lambda_n]$.

- Conjugate gradient iteration, therefore, will generally outperform the estimates given below.

- The estimates nonetheless tend to be quite accurate in practice, however, because the discrete eigenvalues are relatively densely packed on $[\lambda_1, \lambda_n]$.

- The standard Chebyshev polynomials, $T_k(x) = \cos(k \cos^{-1} x)$ have the property that their k roots on the interval $x \in [-1, 1]$ are chosen such that all of their extrema on that interval are the same.

- Here, we are interested in minimizing on the interval $[\lambda_1, \lambda_n]$, subject to $p(0) = 1$.

- Because P_1^k may be any polynomial of degree k satisfying $P_1^k(0) = 1$, we are at liberty to choose one that has the minimal value of M.

- This is given by the scaled and translated Chebyshev polynomial,

$$\tilde{T}_k(\lambda) \;=\; M T_k \left(1 \;-\; 2 \, \frac{\lambda - \lambda_1}{\lambda_n - \lambda_1} \right) \qquad . \tag{22}$$

- Since $T_k(x)$ has extrema ± 1 on the interval $-1 \le x \le 1$, clearly $\tilde{T}_k(\lambda)$ has extrema $\pm M$ on the interval $\lambda_1 \le \lambda \le \lambda_n$.
- From the required scaling, $\tilde{T}_k(0) = 1$, we find

$$M^{-1} \;=\; T_k\left(1 \;-\; 2\,\frac{0-\lambda_1}{\lambda_n - \lambda_1}\right) \;=\; T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right) \;=\; T_k\left(\frac{\kappa+1}{\kappa-1}\right), \quad (23)$$

  where $\kappa = \lambda_n / \lambda_1$.
- It merely remains to evaluate $T_k(x)$ with the appropriate argument to establish the bound.

- We do not go through all of the steps here, but note that the process starts with a representation for the Chebyshev polynomials when the argument of T_k has modulus > 1,

$$T_k(x) \;=\; \frac{1}{2}\left[x + \sqrt{x^2-1}\right]^k \;+\; \frac{1}{2}\left[x - \sqrt{x^2-1}\right]^k . \qquad (24)$$

- After a few pages of manipulation, the desired bound is[1]

$$M \;\leq\; 2\left(\frac{\sqrt{\frac{\lambda_n}{\lambda_1}} - 1}{\sqrt{\frac{\lambda_n}{\lambda_1}} + 1}\right)^k \;=\; 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \quad \text{(CG bound).} \qquad (25)$$

- If $\kappa \gg 1$, then the number of iterations scales as $\sqrt{\kappa}$. With a good preconditioner, however, one can often converge in just a few (e.g., 5–20) iterations.
- The bound (25) is to be contrasted with that for optimal Richardson iteration and steepest descent, both of which have an error bound of the form [Saad],

$$\frac{\|\underline{e}_k\|_A}{\|\underline{x}\|_A} \;\leq\; \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \quad \text{(Richardson/steepest-descent bound).} \qquad (26)$$

- Thus, if either of these methods takes 100 iterations, we can expect CG to take ≈ 10 iterations.

[1] See Saad, *Iterative Methods for Linear Systems*

Deriving the Bound

- We present a sketch of the derivation here.
- The Taylor series arguments are formally correct but the results are more precise than they would indicate, as we mention below.
- From (23) and (24), we have

$$M = \frac{2}{(a+b)^k + (a-b)^k} \leq \frac{2}{(a+b)^k}, \tag{27}$$

where

$$a = \frac{\kappa+1}{\kappa-1}, \qquad b = \sqrt{a^2-1}. \tag{28}$$

- The inequality (27) will generally be quite sharp as k increases because $(a-b)$ will be small compared to $(a+b)$.

- Define $\epsilon := \kappa^{-1} < 1$ and compute the Taylor series expansion for a and b in terms of ϵ,

$$
\begin{aligned}
a &= \frac{\kappa + 1}{\kappa - 1} = \frac{1+\epsilon}{1-\epsilon} && (29)\\
&= (1+\epsilon)(1+\epsilon+\epsilon^2+\dots) \\
&= 1 + 2\epsilon + 2\epsilon^2 + \dots && (30)
\end{aligned}
$$

$$
\begin{aligned}
b &= \left(a^2 - 1\right)^{\frac{1}{2}} && (31)\\
&= \left(1 + 4\epsilon + 8\epsilon^2 + \dots - 1\right)^{\frac{1}{2}} && (32)\\
&= \left(4\epsilon + 8\epsilon^2 + \dots\right)^{\frac{1}{2}} && (33)\\
&= 2\sqrt{\epsilon}\,(1 + \epsilon + \dots)\,. && (34)
\end{aligned}
$$

- Summing a and b and ordering the terms in powers of $\epsilon^{\frac{1}{2}}$, we have

$$
\begin{aligned}
a + b &= 1 + 2\sqrt{\epsilon} + 2\epsilon + 2\epsilon^{\frac{3}{2}} + 2\epsilon^2 + \ldots & (35)\\
&= (1 + \sqrt{\epsilon})(1 + \sqrt{\epsilon} + \epsilon + \epsilon^{\frac{3}{2}} + \ldots) & (36)\\
&\sim \frac{1 + \sqrt{\epsilon}}{1 - \sqrt{\epsilon}}. & (37)
\end{aligned}
$$

- From the preceding result and (27) we have

$$
M \le 2\left(\frac{1 - \sqrt{\epsilon}}{1 + \sqrt{\epsilon}}\right)^k = 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k. \quad (38)
$$

- Note that the Taylor expansions used here would only indicate an asymptotic equivalence ("$\sim$"), but the expressions on the right of (27) and (38) are in fact equal.

Preconditioned Case

- In the unpreconditioned case, our search space was

$$\mathcal{R}(P_k) \;=\; \mathcal{R}(R_k) \;=\; K_k(A;\underline{b}) \;:=\; \mathbb{P}_{k-1}(A)\underline{b} \tag{39}$$

$$P_k \;=\; [\underline{p}_1\ \underline{p}_2\ \dots\ \underline{p}_k] \tag{40}$$

$$R_k \;=\; [\underline{r}_0\ \underline{r}_1\ \dots\ \underline{r}_{k-1}] \tag{41}$$

$$K_k \;=\; span\{\underline{b}\ A\underline{b}\ \dots\ A^{k-1}\underline{b}\}. \tag{42}$$

- Here, we consider a *preconditioned* search space, SPD preconditioner M, $\mathcal{R}(P_k) = \mathcal{R}(Z_k)$, with

$$\underline{z}_k \;:=\; M^{-1}\underline{r}_{k-1} \tag{43}$$

$$\underline{p}_k \;:=\; \underline{z}_k \;-\; \sum_{j=1}^{k-1} \beta_j \underline{p}_j\ . \tag{44}$$

- As before, we will want $\underline{p}_i^T A \underline{p}_j = 0$ for $i \neq j$, so that we find the best-fit (i.e., the projection) with a short term recurrence for $\underline{x}$ and $\underline{r}$.

- Let's look at the Krylov subspace generated by (43)–(44).

- The first few terms in the algorithm yield

$$
\begin{aligned}
\underline{r}_0 &= \underline{b} && \in \mathbb{P}_0(AM^{-1})\underline{b} \\
\underline{p}_1 &= M^{-1}\underline{b} && \in \mathbb{P}_0(M^{-1}A)M^{-1}\underline{b} \\
\underline{r}_1 &= \underline{b} - A\underline{p}_1 && \in \mathbb{P}_1(AM^{-1})\underline{b} \\
\underline{p}_2 &= M^{-1}\left(\underline{b} - A\underline{p}_1\right) - \beta_1 M^{-1}\underline{b} && \in \mathbb{P}_1(M^{-1}A)M^{-1}\underline{b} \\
\underline{r}_2 &= \underline{b} - A\underline{p}_2 && \\
&= \underline{b} - AM^{-1}\left[\underline{b} - A\underline{p}_1 - \beta_1 M^{-1}\right] && \in \mathbb{P}_2(AM^{-1})\underline{b} \\
\underline{p}_3 &= M^{-1}\underline{r}_2 - \sum \beta_j \underline{p}_j && \in \mathbb{P}_2(M^{-1}A)M^{-1}\underline{b}
\end{aligned}
\tag{45}
$$

- Or, in general,

$$\underline{p}_j \;\in\; \mathbb{P}_{j-1}(M^{-1}A)M^{-1}\underline{b} \tag{46}$$

$$A\underline{p}_j \;\in\; \mathbb{P}_j(AM^{-1})\underline{b} \;=\; span\{\underline{r}_0\;\underline{r}_1\;\ldots\underline{r}_j\} \tag{47}$$

- To maintain A-conjugancy, the β_js need to enforce the projection,

$$\underline{p}_k \;:=\; \underline{z}_k \;-\; \Pi_A(P_{k-1})\underline{z}_k \tag{48}$$

$$:=\; \underline{z}_k \;-\; \sum_{j=1}^{k-1}\beta_j\underline{p}_j\;, \tag{49}$$

  with

$$\beta_j \;=\; \frac{\underline{p}_j^T A\underline{z}_k}{\underline{p}_j^T A\underline{p}_j}. \tag{50}$$

- To evaluate the inner-product in the numerator, note

$$A\underline{p}_j \in \mathbb{P}_j(AM^{-1})\underline{b} \tag{51}$$

$$\underline{z}_k = M^{-1}\underline{r}_{k-1} \tag{52}$$

which leads to:

$$\underline{z}_k^T A\underline{p}_j = \underline{r}_{k-1}^T M^{-1} A\underline{p}_j \tag{53}$$

$$M^{-1}A\underline{p}_j \in \mathbb{P}_j(M^{-1}A)M^{-1}\underline{b} \tag{54}$$

$$= span\{\underline{p}_1 \ldots \underline{p}_{j+1}\}. \tag{55}$$

$$M^{-1}A\underline{p}_{k-2} \in span\{\underline{p}_1 \ldots \underline{p}_{k-1}\}. \tag{56}$$

- However, we have

$$\underline{r}_{k-1} \perp_2 \underline{p}_j, \quad j = 1, \ldots, k-1. \tag{57}$$

- So, $\beta_j = 0$ for $j = 1, \ldots, k-2$, and we have only to retain the last term in the series.

- As a result, we again have a short-term recurrence

$$\underline{z} \; := \; M^{-1}\underline{r} \tag{58}$$

$$\underline{p} \; := \; \underline{z} \; - \; \beta\underline{p} \; , \tag{59}$$

with

$$\beta \; = \; \frac{\underline{w}^T\underline{z}}{\underline{w}^T\underline{p}} \tag{60}$$