Projection-Based Iterative Methods, II

Convergence of CG

The CG convergence analysis proceeds from the following observations.

- The kth iterate, $\underline{x}_k$ is the best possible approximation in $K_k(A;\underline{b}) = span\{\underline{b}\ A\underline{b}\ldots\ A^{k-1}\underline{b}\}$
- $\underline{x}_k$ can be expressed as a polynomial in A as $\underline{x}_k = c_1\underline{b} + c_2A\underline{b} + \cdots + c_kA^{k-1}\underline{b} = P_{CG}^{k-1}(A)\underline{b}$, 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

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

To analyze the convergence behavior, notice that the error,

$$\underline{e}_k = \underline{x} - \underline{x}_k = A^{-1}(\underline{b} - A\underline{x}_k) = A^{-1}\underline{r}_k. \tag{2}$$

Thus the A-norm of the error, which is minimized by our approximation, is given by:

$$\begin{aligned} \|\underline{e}_k\|_A^2 &= \underline{e}_k^T A\underline{e}_k \\ &= \underline{r}_k^T A^{-1}\underline{r}_k\ =\ \|\underline{r}_k\|_{A^{-1}}^2 \quad . \end{aligned} \tag{3}$$

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

$$\begin{aligned} \underline{r}_k &= \underline{b} - A\underline{x}_k \\ &= \underline{b} - c_1A\underline{b} - c_2A^2\underline{b} - \ldots - c_kA^k\underline{b} \quad . \end{aligned} \tag{4}$$

Note that the degrees of freedom in (4) are represented by the c_j's. Thus, out of all possible polynomials having the form $P_1^k(t) = 1 + \gamma_1 t + \ldots + \gamma_k t^k$ (i.e., those satisfying $P_1^k(0) = 1$), the conjugate gradient algorithm constructs the one which minimizes $\|\underline{e}_k\|^2$:

$$\begin{aligned} \|\underline{e}_k\|_A^2 &= \underline{r}_k^T A^{-1}\underline{r}_k \\ &= \underline{b}^T(I - AP_{CG}^{k-1})^T A^{-1}(I - AP_{CG}^{k-1})\underline{b} \\ &\leq \underline{b}^T[P_1^k(A)]^T A^{-1}P_1^k(A)\underline{b} \quad , \end{aligned} \tag{5}$$

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

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 $\leq$ 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.

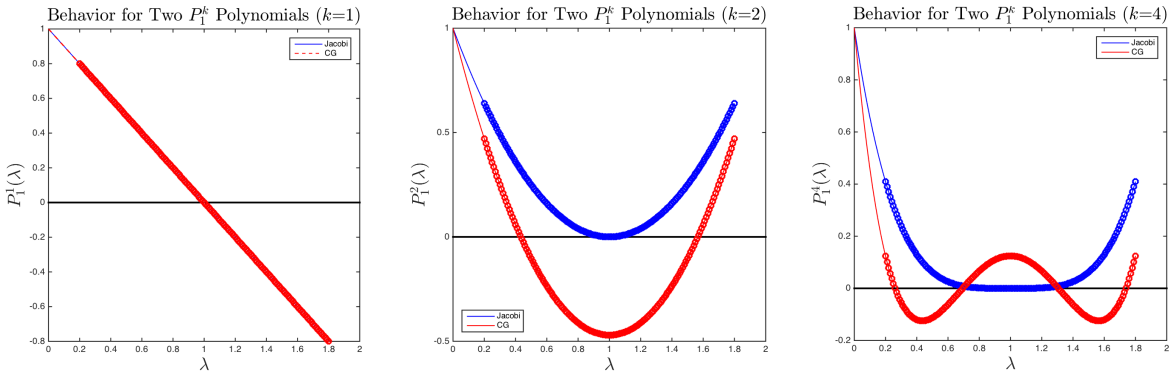

Figure 1: Comparison of error distribution for $\lambda_j \in [0.2{:}1.8]$ for error polynomials based on Jacobi iteration vs. Chebyshev distribution. The CG error distribution will be smaller than the Chebyshev one. (Why?)

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

$$\underline{b} \;=\; \sum_{i=1}^{n} \hat{b}_i \underline{z}_i \,, \tag{6}$$

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

$$\underline{z}_i^T \underline{z}_j \;=\; \delta_{ij} \,, \tag{7}$$

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 (6). The (arbitrary) scaling of the eigenvectors is established by (7). We will use the following relationship shortly.

$$\|\underline{x}\|_A^2 \;=\; \|A^{-1}\underline{b}\|_A^2 \;=\; (A^{-1}\underline{b})^T A (A^{-1}\underline{b}) \;=\; \underline{b}^T A^{-1} \underline{b} \;=\; \sum_{i=1}^{n} \frac{\hat{b}_i^2}{\lambda_i}. \tag{8}$$

Inserting the spectral decomposition (6) of $\underline{b}$ into the error equation (5) yields

$$\|\underline{e}_k\|_A^2 \;\leq\; \left(\sum_{i=1}^{n} P_1^k(\lambda_i) \hat{b}_i \underline{z}_i \right)^T \left(\sum_{j=1}^{n} P_1^k(\lambda_j) \frac{\hat{b}_j}{\lambda_j} \underline{z}_j \right) \tag{9}$$

$$\;=\; \left(\sum_{j=1}^{n} \sum_{i=1}^{n} P_1^k(\lambda_i) P_1^k(\lambda_j) \frac{\hat{b}_i \, \hat{b}_j}{\lambda_j} \underline{z}_i^T \underline{z}_j \right). \tag{10}$$

From the orthonormality of the eigenvectors (7) we have:

$$\|\underline{e}_k\|_A^2 \;\leq\; \sum_{i=1}^{n} (P_1^k(\lambda_i))^2 \frac{\hat{b}_i^2}{\lambda_i} \;\leq\; \sum_{i=1}^{n} M^2 \frac{\hat{b}_i^2}{\lambda_i} \;=\; M^2 \sum_{i=1}^{n} \frac{\hat{b}_i^2}{\lambda_i} \;=\; M^2 \, \|\underline{x}\|_A^2. \tag{11}$$

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

$$M \;:=\; \max_i |P_1^k(\lambda_i)|, \tag{12}$$

which is the bound we seek. We have

$$\frac{\|\underline{e}_k\|}{\|\underline{x}\|} \;\leq\; M \;=\; \max_i |P_1^k(\lambda_i)| \tag{13}$$

$$\;\leq\; \max_{\lambda_1 \leq \lambda \leq \lambda_n} |P_1^k(\lambda)|. \tag{14}$$

Since P_1^k may be *any* polynomial of degree k satisfying $P_1^k(0) = 1$ we can estimate a relatively sharp bound bound by finding a polynomial that minimizes the right-hand side of (14). That is, find

$$P_1^k(\lambda) \;=\; \underset{p\in\mathbb{P}_k^1}{\operatorname{argmin}} \max_{\lambda\in[\lambda_1:\lambda_n]} |p(\lambda)| \tag{15}$$

The solution to this problem, as is often the case in minimax problems, is given by a scaled and translated Chebyshev polynomial, as mentioned previously and discussed further below. Before proceeding with that analysis, however, we note that (13) 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 (9), 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, $\{\, \lambda_1 < \lambda_2 < \ldots < \ldots \lambda_m \,\}$. In this case, $\underline{b}$ has an equivalent spectral decomposition

$$\underline{b} = \sum_{i=1}^{m} \hat{b}_i \underline{z}_i \,, \tag{16}$$

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 (14). This is a classic minimax problem which is invariably solved by using Chebyshev polynomials, $T_k(x)$. We reiterate that (13) provides a *tighter* error bound because the maximum in (13) 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. It is straightforward to show that this also implies that, among all polynomials of degree k satisfying $p(x = 1) = 1$, the Chebyshev polynomials minimize $\max_{x\in[-1,1]} |p(x)|$. 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,

$$\tilde{T}_k(\lambda) \;=\; M T_k\left(1 \,-\, 2\,\frac{\lambda - \lambda_1}{\lambda_n - \lambda_1}\right) \qquad . \tag{17}$$

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

$$M^{-1} \;=\; T_k\left(1 \,-\, 2\,\frac{0 - \lambda_1}{\lambda_n - \lambda_1}\right) \;=\; T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right) \;=\; T_k\left(\frac{\kappa + 1}{\kappa - 1}\right), \tag{18}$$

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,

$$T_k(x) \;=\; \frac{1}{2}\left[x + \sqrt{x^2 - 1}\right]^k \;+\; \frac{1}{2}\left[x - \sqrt{x^2 - 1}\right]^k. \tag{19}$$

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

$$M \;\le\; 2\left(\frac{\sqrt{\frac{\lambda_n}{\lambda_1}} - 1}{\sqrt{\frac{\lambda_n}{\lambda_1}} + 1}\right)^k \;=\; 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \quad \text{(conjugate-gradient bound)} \tag{20}$$

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.

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

The bound (20) is to be contrasted with that for optimal Richardson iteration and steepest descent, both of which have an error bound of the form [Saad],

$$\frac{\|\underline{e}_k\|_A}{\|\underline{x}\|_A} \;\leq\; \left(\frac{\kappa-1}{\kappa+1}\right)^k \quad \text{(Richardson/steepest-descent bound).} \tag{21}$$

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

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 (18) and (19), we have

$$M \;=\; \frac{2}{(a+b)^k + (a-b)^k} \;\leq\; \frac{2}{(a+b)^k}, \tag{22}$$

where

$$a \;=\; \frac{\kappa+1}{\kappa-1} \tag{23}$$

and $b = \sqrt{a^2-1}$. The inequality (22) 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,

$$\begin{aligned} a \;&=\; \frac{\kappa+1}{\kappa-1} \;=\; \frac{1+\epsilon}{1-\epsilon} && (24)\\ &=\; (1+\epsilon)(1+\epsilon+\epsilon^2+\dots) \\ &=\; 1+2\epsilon+2\epsilon^2+\dots && (25) \end{aligned}$$

$$\begin{aligned} b \;&=\; \left(a^2-1\right)^{\frac{1}{2}} && (26)\\ &=\; \left(1+4\epsilon+8\epsilon^2+\cdots-1\right)^{\frac{1}{2}} && (27)\\ &=\; \left(4\epsilon+8\epsilon^2+\dots\right)^{\frac{1}{2}} && (28)\\ &=\; 2\sqrt{\epsilon}\,(1+\epsilon+\dots)\,. && (29) \end{aligned}$$

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

$$\begin{aligned} a+b &= 1+2\sqrt{\epsilon}+2\epsilon+2\epsilon^{\frac{3}{2}}+2\epsilon^2+\dots && (30)\\ &= (1+\sqrt{\epsilon})(1+\sqrt{\epsilon}+\epsilon+\epsilon^{\frac{3}{2}}+\dots) && (31)\\ &\sim \frac{1+\sqrt{\epsilon}}{1-\sqrt{\epsilon}}. && (32) \end{aligned}$$

From the preceding result and (22) we have

$$M \;\leq\; 2\left(\frac{1-\sqrt{\epsilon}}{1+\sqrt{\epsilon}}\right)^k \;=\; 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k. \qquad (33)$$

Note that the Taylor expansions used here would only indicate an asymptotic equivalence ("$\sim$"), but the expressions on the right of (22) and (33) are in fact equal.

Eigenvalues from PCG

Consider the following PCG algorithm with A and preconditioner M SPD.

$$\text{for } k = 1 : n \tag{34}$$

$$\text{Solve } M\underline{z}_k = \underline{r}_{k-1},\ \rho_k = \underline{z}_k^T \underline{r}_{k-1},\ \beta_k = \tfrac{\rho_k}{\rho_{k-1}},\ \text{if } k{=}1{:}\ \beta_k = 0 \tag{35}$$

$$\underline{p}_k \ = \ \underline{z}_k \ + \ \beta_k \underline{p}_{k-1} \tag{36}$$

$$\underline{w}_k \ = \ A\underline{p}_k, \ \ \gamma_k = \underline{p}_k^T \underline{w}_k, \ \ \alpha_k = \frac{\rho_k}{\gamma_k} \tag{37}$$

$$\underline{x}_k \ = \ \underline{x}_{k-1} \ + \ \alpha_k \underline{p}_k \tag{38}$$

$$\underline{r}_k \ = \ \underline{r}_{k-1} \ - \ \alpha_k \underline{w}_k \tag{39}$$

$$\text{end} \tag{40}$$

Let

$$R \ = \ [\underline{r}_0\ \underline{r}_1\ \ldots\ \underline{r}_{k-1}], \tag{41}$$

$$P \ = \ [\underline{p}_1\ \underline{p}_2\ \ldots\ \underline{p}_k], \tag{42}$$

$$Z \ = \ [\underline{z}_1\ \underline{z}_2\ \ldots\ \underline{z}_k], \tag{43}$$

and

$$B \ = \ \begin{bmatrix} 1 & -\beta_2 & & \\ & 1 & -\beta_3 & \\ & & 1 & -\beta_4 \\ & & & \ddots \end{bmatrix}. \tag{44}$$

Note that $P^T AP$ =diag(γ_i). We also have that $Z = PB$, so

$$Z^T AZ \ = \ \tilde{T}, \tag{45}$$

where $\tilde{T}$ is a $k \times k$ tridiagonal matrix. Moreover, $Z^T MZ = Z^T R$ =diag(ρ_i) =: Δ^2. From these, we can find approximate eigenvectors and eigenvalues for the generalized eigenvalue problem,

$$A\underline{s}_j \ = \ \lambda_j M\underline{s}_j. \tag{46}$$

Let's consider λ_n, the eigenvalue that maximizes the Rayleigh quotient,

$$\lambda_n \ = \ \max_{\underline{s} \in \mathbb{R}^n} \frac{\underline{s}^T A\underline{s}}{\underline{s}^T M\underline{s}}. \tag{47}$$

The approximated value is

$$\lambda_n \approx \max_{\underline{z}\in Z} \frac{\underline{z}^T A \underline{z}}{\underline{z}^T M \underline{z}} = \max_{\underline{y}\in \mathbb{R}^k} \frac{\underline{y}^T Z^T A Z \underline{y}}{\underline{y}^T Z^T M Z \underline{y}} = \max_{\underline{y}\in \mathbb{R}^k} \frac{\underline{y}^T \tilde{T} \underline{y}}{\underline{y}^T \Delta^2 \underline{y}} = \mu_k, \tag{48}$$

where μ_k is the maximum eigenvalue for the $k \times k$ eigenvalue problem,

$$\tilde{T}\underline{y}_j = \mu_j \Delta^2 \underline{y}_j \tag{49}$$

Here, Δ^2 is a diagonal matrix. Define $\underline{u} = \Delta \underline{y} \longrightarrow \underline{y} = \Delta^{-1}\underline{u}$. The right-most Rayleigh quotient in the preceding equation becomes

$$\mu_k = = \max_{\underline{u}\in \mathbb{R}^k} \frac{\underline{u}^T T \underline{u}}{\underline{u}^T \underline{u}}, \tag{50}$$

where $T := \Delta^{-1}\tilde{T}\Delta^{-1}$.

As a result, PCG is effectively a Lanczos algorithm from which one can approximate the (extreme) eigenvalues and eigenvectors of (46). The approximations are given by μ_j and $Z\underline{y}_j$ from solving the small $k \times k$ eigenvalue problem (49).