Numerical Analysis - Part II

Anders C. Hansen

Lecture 20

Iterative methods for linear algebraic systems

The conjugate gradient method

Here it is.

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

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

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

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

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

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

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

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

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

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

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

Proof. We use induction on $m \geq 0$, the assertions being trivial for $m = 0$, since $\boldsymbol{d}^{(0)} = \boldsymbol{r}^{(0)}$ 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)

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

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

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

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

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

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

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

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

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

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

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

By the inductive hypothesis $\boldsymbol{d}^{(k)} \perp A\boldsymbol{d}^{(i)}$, hence it remains to establish that $\boldsymbol{r}^{(k+1)} \perp A\boldsymbol{d}^{(i)}$ for $i < k$. Now, the formula (1) yields $A\boldsymbol{d}^{(i)} = (\boldsymbol{r}^{(i)} - \boldsymbol{r}^{(i+1)})/\alpha_i$, therefore we require the conditions $\boldsymbol{r}^{(k+1)} \perp (\boldsymbol{r}^{(i)} - \boldsymbol{r}^{(i+1)})$ 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 $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$*, termination occurs after at most* n *iterations. More precisely, termination occurs after at most* s *iterations, where* $s = \dim \operatorname{span}\{A^i \boldsymbol{r}_0\}_{i=0}^{n-1}$ *(which can be smaller than* n*).*

Termination property

Proof. Assertion (2) of Theorem 1 states that residuals $(\boldsymbol{r}^{(k)})_{k\geq 0}$ 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 $\operatorname{span}\{A^i \boldsymbol{r}_0\}_{i=0}^{n-1}$, their number is bounded by the dimension of that space. □

Definition 3 (The Krylov subspaces)

Let A be an $n \times n$ matrix, $\boldsymbol{v} \in \mathbb{R}^n$ nonzero, and $m \in \mathbb{N}$. The linear space $K_m(A, \boldsymbol{v}) := \operatorname{span}\{A^i \boldsymbol{v}\}_{i=0}^{m-1}$ 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 $\boldsymbol{v}$,

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

Hence, for any $A > 0$, the number of iterations of the CGM for solving $A\boldsymbol{x} = \boldsymbol{b}$ 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 $(\boldsymbol{u}_i)$. Indeed, in that case one can expand $\boldsymbol{v} = \sum_{i=1}^{n} a_i \boldsymbol{u}_i$, and then group together eigenvectors with the same eigenvalues: for each λ_ν we set $\boldsymbol{w}_\nu = \sum_{k=1}^{m_\nu} a_{i_k} \boldsymbol{u}_{i_k}$ if $A\boldsymbol{u}_{i_k} = \lambda_\nu \boldsymbol{u}_{i_k}$. Then

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

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

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

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

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

Standard form of the conjugate gradient method

Here it is.

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

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

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

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

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

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

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

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

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

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,

`https://en.wikipedia.org/wiki/Cholesky_decomposition`

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

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

for some $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$, but note that computing $Q^{-1}\boldsymbol{z}$ is the same as solving the linear system $Q\boldsymbol{v} = \boldsymbol{z}$, 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 $A\boldsymbol{x} = \boldsymbol{b}$ 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 $A = S + \boldsymbol{e}_1 \boldsymbol{e}_1^T$, we have

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

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

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

where $\boldsymbol{e}^{(k)} = x^* - x^{(k)}$ *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* $\{\boldsymbol{d}^{(k)}\}_{k=0}^{m-1}$ *be a set of the conjugate directions, i.e.,* $(A\boldsymbol{d}^{(k)}, \boldsymbol{d}^{(i)}) = 0$ *for* $i < k$, *and consider*

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

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

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

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

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

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

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

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 $\boldsymbol{x}^{(k)}$ that minimizes the functional

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

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

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

Hence, because of the minimization property of CGM,

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

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:

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

□

Eigenvalues and eigenvectors

Motivation – The Schrödinger equation

One of the word's most famous eigenvalue problems: The time independent Schrödinger equation

$$\left[\frac{-\hbar^2}{2m}\nabla^2 + V(\mathrm{r})\right]\Psi(\mathrm{r}) = E\Psi(\mathrm{r}).$$

Introduction to matrix eigenvalue calculations

Let A be a real $n \times n$ matrix. The eigenvalue equation is $A\boldsymbol{w} = \lambda \boldsymbol{w}$, where λ is a scalar, which may be complex if A is not symmetric. There exists a nonzero vector $\boldsymbol{w} \in \mathbb{R}^n$ satisfying this equation if and only if $\det(A - \lambda I) = 0$. The function $p(\lambda) = \det(A - \lambda I)$, $\lambda \in \mathbb{C}$, is a polynomial of degree n, but calculating the eigenvalues by finding the roots of p is a disaster area because of loss of accuracy due to rounding errors.

If the polynomial has some multiple roots and if A is not symmetric, then the number of linearly independent eigenvectors may be fewer than n, but there are always n mutually orthogonal real eigenvectors in the symmetric case.

We assume in all cases, however, that the eigenvalue equations $A\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$, $i = 1..n$, are satisfied by eigenvectors $\boldsymbol{w}_i$ that are linearly independent, which can be achieved by making an arbitrarily small change to A if necessary.

The power method

The iterative algorithms that will be studied for the calculation of eigenvalues and eigenvectors are all closely related to the power method, which has the following basic form for generating a single eigenvalue and eigenvector of A.

We pick a nonzero vector $\boldsymbol{x}^{(0)}$ in $\mathbb{R}^n$. Then, for $k = 0, 1, 2, \ldots$, we let $\boldsymbol{x}^{(k+1)}$ be a nonzero multiple of $A\boldsymbol{x}^{(k)}$, typically to satisfy $\|\boldsymbol{x}^{(k+1)}\| = 1$ so that

$$\boldsymbol{x}^{(k+1)} = A\boldsymbol{x}^{(k)}/\|A\boldsymbol{x}^{(k)}\|, \qquad k = 0, 1, 2, \ldots$$

This method is oriented on finding an eigenvector corresponding to the largest eigenvalue as the the following theorem shows.