Numerical Linear Algebra

Maya Neytcheva, TDB, February-March 2021

Projection-type methods continued

- Recall: Arnoldi/Lanczos process
- The Conjugate Gradient method - derivation, properties and convergence
- The GMRES method - derivation, properties and convergence
- Optional: The Generalized Conjugate Gradient method - derivation, properties and convergence

How to construct a basis for $\mathcal{K}$?

Arnoldi's method for general matrices

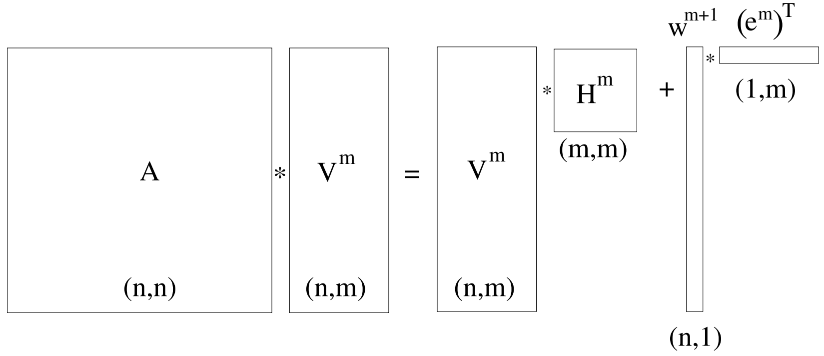
Consider $\mathcal{K}^m(A, \mathbf{v}) = \{\mathbf{v}, A\mathbf{v}, A^2\mathbf{v}, \cdots, A^{m-1}\mathbf{v}\}$, generated by some matrix A and a vector $\mathbf{v}$.

1. Choose a vector $\mathbf{v}^{(1)}$ such that $\|\mathbf{v}^{(1)}\| = 1$
2. For $k = 1, 2, \cdots, m$
3. For $i = 1, 2, \cdots, k$
4. $h_{ik} = (A\mathbf{v}^{(k)}, \mathbf{v}^{(i)})$
5. End
6. $\mathbf{w}^{(k)} = A\mathbf{v}^{(k)} - \sum_{i=1}^{k} h_{ik}\mathbf{v}^{(i)}$
7. $h_{k+1,k} = \|\mathbf{w}^{(k)}\|$
8. If $h_{k+1,k} = 0$, stop
9. $\mathbf{v}^{(k+1)} = \mathbf{w}^{(k)}/h_{k+1,k}$
10. End

Memory demands: we keep all vectors $\mathbf{v}^{(k)}$ and $A\mathbf{v}^{(k)}$, $k = 1, \cdots, m$.

The result of Arnoldi's process

- $V^m = \{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \cdots, \mathbf{v}^{(m)}\}$ is an orthonormal basis in $\mathcal{K}^m(A, \mathbf{v})$
- $AV^m = V^m H^m + \mathbf{w}^{m+1}\mathbf{e}_m^T$



Arnoldi's process - example

$$H^{(3)} = \begin{bmatrix} (A\mathbf{v}^{(1)}, \mathbf{v}^{(1)}) & (A\mathbf{v}^{(2)}, \mathbf{v}^{(1)}) & (A\mathbf{v}^{(3)}, \mathbf{v}^{(1)}) \\ \|\mathbf{w}^1\| & (A\mathbf{v}^{(2)}, \mathbf{v}^{(2)}) & (A\mathbf{v}^{(3)}, \mathbf{v}^{(2)}) \\ 0 & \|\mathbf{w}^{(2)}\| & (A\mathbf{v}^{(3)}, \mathbf{v}^{(3)}) \end{bmatrix}$$

Since $V^{m+1} \perp \{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \cdots, \mathbf{v}^m\}$ then it follows that $(V^m)^T A V^m = H^m$.
H^m is an upper-Hessenberg matrix.

$$\overline{H}^{(3)} = \begin{bmatrix} (A\mathbf{v}^{(1)}, \mathbf{v}^{(1)}) & (A\mathbf{v}^{(2)}, \mathbf{v}^{(1)}) & (A\mathbf{v}^{(3)}, \mathbf{v}^{(1)}) \\ \|\mathbf{w}^1\| & (A\mathbf{v}^{(2)}, \mathbf{v}^{(2)}) & (A\mathbf{v}^{(3)}, \mathbf{v}^{(2)}) \\ 0 & \|\mathbf{w}^{(2)}\| & (A\mathbf{v}^{(3)}, \mathbf{v}^{(3)}) \\ 0 & 0 & \|\mathbf{w}^{(3)}\| \end{bmatrix}$$

Arnoldi's method for symmetric matrices

For A - real symmetric, Arnoldi's method reduces to the Lanczos method.

Recall: $H^m = (V^m)^T A V^m$
If A is symmetric, then H^m must be symmetric too, i.e., H^m is three-diagonal

$$H^m = \begin{bmatrix} \gamma_1 & \beta_2 & & \\ \beta_2 & \gamma_2 & \beta_3 & \\ & & \ddots & \\ & & \beta_m & \gamma_m \end{bmatrix}$$

Thus, the vectors $\mathbf{v}^{(k)}$ satisfy a three-term recursion:

$$\beta_{k+1}\mathbf{v}^{(k+1)} = A\mathbf{v}^{(k)} - \gamma_k \mathbf{v}^{(k)} - \beta_k \mathbf{v}^{(k-1)}$$

Arnoldi

1. $\mathbf{v}^{(1)}$ such that $\|\mathbf{v}^{(1)}\| = 1$
2. For $k = 1, 2, \cdots, m$
3. For $i = 1, 2, \cdots, k$
4. $h_{ik} = (A\mathbf{v}^{(k)}, \mathbf{v}^{(i)})$
5. End
6. $\mathbf{w}^{(k)} = A\mathbf{v}^{(k)} - \sum_{i=1}^{(k)} h_{ik}\mathbf{v}^{(i)}$
7. $h_{k+1,k} = \|\mathbf{w}^{(k)}\|$
8. If $h_{k+1,k} = 0$, stop
9. $\mathbf{v}^{(k+1)} = \mathbf{w}^{(k)} / h_{k+1,k}$
10. End

Lanczos

$\mathbf{w}^{(0)}$, $\beta = \|\mathbf{w}^{(0)}\|$, $\mathbf{v}^{(1)} = \mathbf{w}^{(0)}/\beta$
For $k = 1, 2, \cdots, m$

$\mathbf{w}^{(k)} = A\mathbf{v}^{(k)} - \beta_k \mathbf{v}^{(k-1)}$
$\gamma_k = (\mathbf{w}^{(k)}, \mathbf{v}^{(k)})$

$\mathbf{w}^{(k)} = \mathbf{w}^{(k)} - \gamma_k \mathbf{v}^{(k)}$

$\beta_{k+1} = \|\mathbf{w}^{(k)}\|$
if $\beta_{k+1} = 0$, stop
$\mathbf{v}^{(k+1)} = \mathbf{w}^{(k)} / \beta_{k+1}$
End
Set $T_m = tridiag\{\beta_k, \gamma_k, \beta_{k+1}\}$

Lanczos algorithm to solve symmetric linear systems

Given: $\mathbf{x}^{(0)}$
Compute $\mathbf{r}^{(0)} = \mathbf{b} - A\mathbf{x}^{(0)}$, $\beta = \|\mathbf{r}^{(0)}\|$, $\mathbf{v}^{(1)} = \mathbf{r}^{(0)}/\beta$
Set $\beta_1 = 0$ and $\mathbf{v}^{(0)} = \mathbf{0}$
For $k = 1 : m$
  $\mathbf{w}^{(k)} = A\mathbf{v}^{(k)} - \beta_k \mathbf{v}^{(k-1)}$
  $\gamma_k = (\mathbf{w}^{(k)}, \mathbf{v}^{(k)})$
  $\mathbf{w}^{(k)} = \mathbf{w}^{(k)} - \gamma_k \mathbf{v}^{(k)}$
  $\beta_{k+1} = \|\mathbf{w}^{(k)}\|$, if $\beta_{k+1} = 0$, go out of the loop
  $\mathbf{v}^{(k+1)} = \mathbf{w}^{(k)}/\beta_{k+1}$
End
Set $T_m = tridiag\{\beta_k, \gamma_k, \beta_{k+1}\}$
Compute $\mathbf{y}_m = T_m^{-1}(\beta \mathbf{e}^{(1)}) \Longleftarrow \Longleftarrow$
  $\mathbf{x}_m = \mathbf{x}_0 + V_m \mathbf{y}_m$

The CG method

Observations regarding CG: (1)

Relation 1: The residuals are orthogonal to each other.
Proof: We have $\mathbf{r}^{(m)} = \mathbf{b} - A\mathbf{x}^{(m)}$.
Then $\mathbf{b} - A\mathbf{x}^{(m)} = -\beta_{m+1}\mathbf{e}_m\mathbf{y}^{(m)}\mathbf{v}^{(m+1)} = const\,\mathbf{v}^{(m+1)}$.

To see the latter, recall that $\mathbf{y}^m = T_m^{-1}(\beta \mathbf{e}_1)$ and
$\mathbf{x}^{(m)} = \mathbf{x}^{(0)} + V^m\mathbf{y}^{(m)} = \mathbf{x}^{(0)} + V^m T_m^{-1}\beta\mathbf{e}_1$
Then,

$$\begin{aligned}
\mathbf{b} - A\mathbf{x}^{(m)} &= \underbrace{\mathbf{b} - A\mathbf{x}^{(0)}}_{\beta\mathbf{v}^1} - AV^m T_m^{-1}\beta\mathbf{e}_1^T \\
&= \beta\mathbf{v}^1 - (V^m H_m \mathbf{y}^{(m)} + \mathbf{h}_{m+1,m}\mathbf{e}_m^T\mathbf{y}^{(m)}\mathbf{v}^{(m+1)}) \\
&= \underbrace{\beta\mathbf{v}^1 - V^m\beta\mathbf{e}_1^T}_{0} - \underbrace{\mathbf{h}_{m+1,m}\mathbf{e}_m^T\mathbf{y}^{(m)}}_{const}\mathbf{v}^{(m+1)}
\end{aligned}$$

Observations regarding CG: (1), cont.

Thus, $\mathbf{r}^{(m)}$ is collinear with $\mathbf{v}^{(m+1)}$.
Since $\mathbf{v}^j$ are orthogonal to each other, then the residuals are also mutually orthogonal, i.e.,
$(\mathbf{r}^{(k)}, \mathbf{r}^{(m)}) = 0$ for $k \neq m$.

Observations re. CG: (2)

Denote $G = V^m L^{-T}$, $G = \{\mathbf{g}^1, \mathbf{g}^2, \cdots, \mathbf{g}^m\}$.
Relation 2: The vectors $\mathbf{g}^j$ are A-conjugate, i.e., $(A\mathbf{g}^i, \mathbf{g}^j) = 0$ for $i \neq j$.
Proof:

$$(V^m)^T A V^m = T_m = LL^T$$
$$(V^m)^T A \underbrace{V^m L^{-T}}_{G} = L$$
$$\underbrace{G^T A G}_{symmetric} = L^{-T}(V^m)^T A V^m V^m L^{-T} = \underbrace{L^{-T} L}_{lowertriang.}$$

Thus, $L^{-T}L$ must be diagonal.

Derivation of the CG method:

(i) $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \tau_k \mathbf{g}^{(k)}$

$$\begin{aligned}
\mathbf{b} - A\mathbf{x}^{(k+1)} &= \mathbf{b} - A\mathbf{x}^{(k)} - \tau_k A\mathbf{g}^{(k)} \\
\mathbf{r}^{(k+1)} &= \mathbf{r}^{(k)} - \tau_k A\mathbf{g}^{(k)} \\
A\mathbf{g}^{(k)} &= \frac{1}{\tau_k}(\mathbf{r}^{(k)} - \mathbf{r}^{(k+1)}) \\
0 &= (\mathbf{r}^{(k)}, \mathbf{r}^{(k)}) - \tau_k (A\mathbf{g}^{(k)}, \mathbf{r}^{(k)}) \\
\Rightarrow \tau_k &= \frac{(\mathbf{r}^{(k)}, \mathbf{r}^{(k)})}{(A\mathbf{g}^{(k)}, \mathbf{r}^{(k)})}
\end{aligned}$$

Derivation of the CG method, cont.:

(ii) $\mathbf{g}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{g}^{(k)}$ Why is this so?
From the algorithm we have that
$\mathbf{g}^{(k+1)} = c_1 \mathbf{v}^{(k+1)} + c_2 \mathbf{g}^{(k)}$ for some constants c_1, c_2.
We get (ii) after a proper scaling. Then
$(A\mathbf{g}^{(k+1)}, \mathbf{g}^{(k+1)}) = (A\mathbf{g}^{(k+1)}, \mathbf{r}^{(k+1)}) + \beta_k \underbrace{(A\mathbf{g}^{(k+1)}, \mathbf{g}^{(k)})}_{0}$

$$\Rightarrow \tau_k = \frac{(\mathbf{r}^{(k)}, \mathbf{r}^{(k)})}{(A\mathbf{g}^{(k)}, \mathbf{g}^{(k)})}$$
$$\beta_k = \frac{(\mathbf{r}^{(k+1)}, A\mathbf{g}^{(k)})}{(\mathbf{g}^{(k)}, A\mathbf{g}^{(k)})} = \frac{(\mathbf{r}^{(k+1)}, \frac{1}{\tau_k}(\mathbf{r}^{(k)} - \mathbf{r}^{(k+1)}))}{(\mathbf{g}^{(k)}, A\mathbf{g}^{(k)})} = -\frac{(\mathbf{r}^{(k+1)}, \mathbf{r}^{(k+1)})}{(\mathbf{r}^{(k)}, \mathbf{r}^{(k)})}$$

Derivation of the CG method (cont):

Rewrite the CG algorithm using the above relations:

Initialize: $\mathbf{r}^{(0)} = A\mathbf{x}^{(0)} - \mathbf{b}$, $\mathbf{g}^{(0)} = \mathbf{r}^{(0)}$
For $k = 0, 1, \cdots$, until convergence
$\tau_k = \frac{(\mathbf{r}^{(k)}, \mathbf{r}^{(k)})}{(A\mathbf{g}^{(k)}, \mathbf{g}^{(k)})}$
$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \tau_k \mathbf{g}^{(k)}$
$\mathbf{r}^{(k+1)} = \mathbf{r}^{(k)} + \tau_k A\mathbf{g}^{(k)}$
$\beta_k = \frac{(\mathbf{r}^{(k+1)}, \mathbf{r}^{(k+1)})}{(\mathbf{r}^{(k)}, \mathbf{r}^{(k)})}$
$\mathbf{g}^{(k+1)} = \mathbf{r}^{(k+1)} + \beta_k \mathbf{g}^{(k)}$
end

$\mathbf{r}^{(k)}$ – iteratively computed residuals
$\mathbf{g}^{(k)}$ – search directions
Note: the coefficients β_k are different from those in the Lanczos method.

CG: computer implementation

```
x = x0
r = A*x-b
delta0 = (r,r)
g = -r
Repeat: h   = A*g
        tau = delta0/(g,h)
        x   = x + tau*g
        r   = r + tau*h
        delta1 = (r,r)
        if delta1 <= eps, stop
        beta = delta1/delta0
        g    = -r + beta*g
```


Optimality properties of the CG method

Opt1: Mutually orthogonal search directions:
$(\mathbf{g}^{(k+1)}, A\mathbf{g}^j) = 0, j = 0, \cdots, k$

Opt2: There holds $\mathbf{r}^{(k+1)} \perp K_m(A, \mathbf{r}^{(0)})$, i.e.,
$(\mathbf{r}^{(k+1)}, A\mathbf{r}^{(k)}) = 0, j = 0, \cdots, k$

Opt3: Optimization property: $\|\mathbf{r}^{(k)}\|$ smallest possible at any step, since CG minimizes the functional
$f(\mathbf{x}) = 1/2(\mathbf{x}, A\mathbf{x}) - (\mathbf{x}, \mathbf{b})$

Opt4: $(\mathbf{e}^{(k+1),A\mathbf{g}^j}) = (\mathbf{g}^{(k+1)}, A\mathbf{g}^j) = (\mathbf{r}^{(k+1)}, \mathbf{r}^{(k)}) = 0, j = 0, \cdots, k$

Optimality properties of the CG method

Opt5: Finite termination property: there are <u>no</u> breakdowns of the CG algorithm.
Reasoning: if $\mathbf{g}^{(k)} = \mathbf{0}$ then τ_k is not defined. the vectors $\mathbf{g}^{(k)}$ are computed from the formula $\mathbf{g}^{(k)} = \mathbf{r}^{(k)} + \beta_k \mathbf{g}^{k-1}$.
Then

$$0 = (\mathbf{r}^{(k)}, \mathbf{g}^{(k)}) = -(\mathbf{r}^{(k)}, \mathbf{r}^{(k)}) + \beta_k \underbrace{(\mathbf{r}^{(k)}, \mathbf{g}^{k-1})}_{0}, \Rightarrow \ \mathbf{r}^{(k)} = \mathbf{0},$$

i.e., the solution is already found.
As soon as $\mathbf{x}^{(k)} \neq \mathbf{x}_{exact}$, then $\mathbf{r}^{(k)} \neq \mathbf{0}$ and then $\mathbf{g}^{(k+1)}) \neq \mathbf{0}$.
However, we can generate at most n mutually orthogonal vectors in R^n, thus, CG has a finite termination property.

Connection to the matrix T_m

The general form of the m-dimensional Lanczos tri-diagonal matrix T_m in terms of the CG coefficients:

$$T_m = \begin{bmatrix} \frac{1}{\tau_0} & \sqrt{\beta_0} & & \\ \sqrt{\beta_0} & \frac{1}{\tau_1} + \frac{\beta_0}{\tau_0} & \frac{\sqrt{\beta_1}}{\tau_1} & \\ \cdot & \cdot & \cdot & \frac{\sqrt{\beta_{m-2}}}{\tau_{m-2}} \\ & & \frac{\sqrt{\beta_{m-2}}}{\tau_{m-2}} & \frac{1}{\tau_{m-1}} + \frac{\beta_{m-2}}{\tau_{m-2}} \end{bmatrix}$$

$$\alpha_k = \frac{1}{\tau_{k-1}} + \frac{\beta_{k-1}}{\tau_{k-2}}, \ \eta_{k+1} = \frac{\sqrt{\beta_k}}{\tau_{k-1}}, \ \beta_0 = 0, \ \tau_{-1} = 1$$

Demo

Convergence analysis of the CG method

Theorem: In exact arithmetic, CG has the property that $\mathbf{x}_{exact} = \mathbf{x}^{(m)}$ for some $m \leq n$, where n is the order of A.

Let $S = \{\lambda_i, \mathbf{s}^i\}_{i=1}^n$ be the system of eigensolutions of A.

Let $\mathbf{r}^{(0)} = \sum_{i=1}^n \xi_i \mathbf{s}^i$. Then, $\mathbf{g}^{(k)} = p_{k-1}(A)\mathbf{r}^{(0)}$, where $p_{k-1}(t)$ is some polynomial of degree $k-1$.

Note: $\mathbf{e}^k = \mathbf{x}_{exact} - \mathbf{x}^{(k)}$, thus, $A\mathbf{e}^k = \mathbf{b} - A\mathbf{x}^{(k)} = \mathbf{r}^{(k)}$.

$$\boxed{\mathbf{e}^k = A^{-1}\mathbf{r}^{(k)} \qquad (**)}$$

CG is such that $\|\mathbf{e}^k\|_A = \min_{\mathbf{y}\in\mathbf{x}^{(0)}+K} \|\mathbf{x}_{exact} - \mathbf{y}\|_A$

From $(**)$ we obtain $\|\mathbf{e}^k\|_A = \|\mathbf{r}^{(k)}\|_{A^{-1}}$

$$\Rightarrow \quad \|\mathbf{r}^{(k)}\|_{A^{-1}} = \min_{\mathbf{r}\in\mathbf{r}^{(0)}+K} \|\mathbf{r}\|_{A^{-1}}$$

Convergence of the CG method (cont)

Let $\Pi_k^1 = \{P_k \text{ of degree } k, P_k(0) = 1\}$ and $\widetilde{K} = \{\mathbf{r} \in R^m : \mathbf{r} = P_k(\mathbf{r}^{(0)}), P_k \in \Pi_k^1\}$.
Clearly, $\widetilde{K} \subset K^k(A, \mathbf{r}^{(0)})$ and $\mathbf{r}^{(0)} \in \widetilde{K}$. Then

$$\begin{aligned}\|\mathbf{r}^{(k)}\|_{A^{-1}} &= \min_{\mathbf{r}\in\widetilde{K}} \|\mathbf{r}\|_{A^{-1}} \\ &= \min_{P_k\in\Pi_k^1} \|P_k(A)\mathbf{r}^{(0)}\|_{A^{-1}} \\ &= \min_{P_k\in\Pi_k^1} \left((\mathbf{r}^{(0)})^T A^{-1}(P_k(A))^2\mathbf{r}^{(0)}\right)^{1/2}\end{aligned}$$

Recall: $(P_k(A))^T A^{-1} P_k(A) = A^{-1}(P_k(A))^2$.

Rate of convergence of the CG method

Theorem: Let A be symmetric and positive definite.
Suppose that for some set S, containing all eigenvalues of A, for some polynomial $\widetilde{P}(\lambda) \in \Pi_k^1$ and some constant M there holds $\max_{\lambda\in S} \left|\widetilde{P}(\lambda)\right| \leq M$.
Then,

$$\|\mathbf{x}_{exact} - \mathbf{x}^{(k)}\|_A \leq M\|\mathbf{x}_{exact} - \mathbf{x}^{(0)}\|_A.$$

Proof: Let $S = \{\lambda_i, \mathbf{s}^i\}_{i=1}^n$ be the system of eigensolutions of A, $\lambda_1 \leq \cdots \leq \lambda_n$, $(\mathbf{s}^i, \mathbf{s}^j) = \delta_{ij}$.
$\mathbf{r}^{(0)} = \sum_{i=1}^n \xi_i \mathbf{s}^i$, $\xi_i = (\mathbf{s}^i, \mathbf{r}^{(0)})$.
Then,

$$(\mathbf{r}^{(0)})^T A^{-1}(P_k(A))^2\mathbf{r}^{(0)} = \sum_{i=1}^n \xi_i^2 \lambda_i^{-1} P_k(\lambda_i)^2$$

$$\Rightarrow \|\mathbf{r}^{(0)}\|_{A^{-1}} = \min_{P_k\in\Pi_k^1} \sum_{i=1}^n \xi_i^2 \lambda_i^{-1} P_k(\lambda_i)^2$$

$$\Rightarrow \|\mathbf{r}^{(k)}\|_{A^{-1}} \leq M^2 \sum_{i=1}^n \xi_i^2 \lambda_i^{-1} = M^2 \|\mathbf{r}^{(0)}\|_{A^{-1}}$$

Rate of convergence (cont)

To quantify M, we seek a polynomial $\widetilde{P}_k \in \Pi_k^1$, such that

$$M = \max_{\lambda \in S} \left| \widetilde{P}_k(\lambda) \right|$$

is small.
In this way, the convergence estimate is replaced by a polynomial approximation problem, which is well known.
For an s.p.d. matrix A and $I_S = [\lambda_1, \lambda_n]$ find a polynomial $\widetilde{P}_k \in \Pi_k^1$ such that

$$\max_{\lambda \in I_S} \left| \widetilde{P}_k(\lambda) \right| = \min_{P_k \in \Pi_k^1} \max_{\lambda \in I_S} |P_k(\lambda)|$$

Rate of convergence (cont):

Repeat: $\max_{\lambda \in I_S} \left| \widetilde{P}_k(\lambda) \right| = \min_{P_k \in \Pi_k^1} \max_{\lambda \in I_S} |P_k(\lambda)|$
The solution of the latter problem is given by the polynomial

$$\widetilde{P}_k(\lambda) = \frac{T_k\left(\frac{\lambda_n + \lambda_1 - 2\lambda}{\lambda_n - \lambda_1}\right)}{T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)}$$

where $T_k(z) = \frac{1}{2}(z^k + z^{-1})$ are the Chebyshev polynomials of degree k. Moreover,

$$\max_{\lambda \in I_S} |P_k(\lambda)| = \frac{1}{T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)}.$$

Rate of convergence (cont):

Thus, we obtain the following estimate:

$$\|\mathbf{e^k}\|_A \leq \frac{1}{T_k\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)} \|\mathbf{e^0}\|_A = \frac{1}{T_k\left(\frac{\varkappa(A)+1}{\varkappa(A)-1}\right)} \|\mathbf{e^0}\|_A$$

Since for any z, $T_k(\frac{z+1}{z-1}) = \frac{1}{2}\left[\left(\frac{\sqrt{z}+1}{\sqrt{z}-1}\right)^k + \left(\frac{\sqrt{z}-1}{\sqrt{z}+1}\right)^k\right] > \frac{1}{2}\left(\frac{\sqrt{z}+1}{\sqrt{z}-1}\right)^k$,

$$\|\mathbf{e^k}\|_A \leq 2\left[\frac{\sqrt{\varkappa(A)} - 1}{\sqrt{\varkappa(A)} + 1}\right]^k \|\mathbf{e^0}\|_A$$

Rate of convergence (cont)

Repeat:

$$\|\mathbf{e^k}\|_A \leq 2\left[\frac{\varkappa(A) - 1}{\varkappa(A) + 1}\right]^k \|\mathbf{e^0}\|_A$$

Seek now the smallest k, such that

$$\|\mathbf{e}^k\|_A \leq \varepsilon \|\mathbf{e}^0\|_A$$

we want $\left(\frac{\sqrt{\varkappa}+1}{\sqrt{\varkappa}-1}\right)^k > \frac{2}{\varepsilon}$

$\Rightarrow\ k\ ln\left(\frac{\sqrt{\varkappa}+1}{\sqrt{\varkappa}-1}\right) > ln(\frac{2}{\varepsilon})$

$\Rightarrow\ k > ln(\frac{2}{\varepsilon})/ln\left(\frac{\sqrt{\varkappa}+1}{\sqrt{\varkappa}-1}\right) = ln(\frac{2}{\varepsilon})/ln\left(\frac{1+(\sqrt{\varkappa})^{-1}}{1-(\sqrt{\varkappa})^{-1}}\right)$

We are on the safe side if
$k > \frac{1}{2}\sqrt{\varkappa}\ ln(\frac{2}{\varepsilon}) > ln(\frac{2}{\varepsilon})/ln\left(\frac{1+(\sqrt{\varkappa})^{-1}}{1-(\sqrt{\varkappa})^{-1}}\right)$

Note: $ln(\frac{1+\epsilon}{1-\epsilon}) > 2\epsilon$ for small ϵ.

Alternative view-point

Let $f(\mathbf{x})$ be a vector function and we restrict $\mathbf{x}$ to be of the form $\mathbf{x} = \mathbf{x} + \tau\mathbf{d}$.
We pose the problem to minimize $f(\mathbf{x})$ for such choice of $\mathbf{x}$.
Since $\mathbf{x}^* + \tau\mathbf{d}$ is a *line*, $\mathbf{d}$ is called a *search direction* and the process is called *line search*.
Consider the special vector function $f^*(\mathbf{x}) = (\mathbf{x}^* - \mathbf{x}, A(\mathbf{x}^* - \mathbf{x}))$. The minimum of $f^*(\mathbf{x})$ coincides with the minimum of $f(\mathbf{x}) = f^*(\mathbf{x}) + C$, where C is constant. For instance, we can take
$C = -\frac{1}{2}(\mathbf{b}, \mathbf{x}^*) + c_0$. Then

$$\begin{aligned} f(\mathbf{x}) &= \tfrac{1}{2} f^*(\mathbf{x}) - \tfrac{1}{2}(\mathbf{b}, \mathbf{x}^*) + c_0 \\ &= \tfrac{1}{2}(\mathbf{x}^* - \mathbf{x}, A(\mathbf{x}^* - \mathbf{x})) - \tfrac{1}{2}(\mathbf{b}, \mathbf{x}^*) + c_0 \\ &= \tfrac{1}{2}(\mathbf{x}^*, A\mathbf{x}^*) - \tfrac{2}{2}(\mathbf{x}, A\mathbf{x}^*) + \tfrac{1}{2}(\mathbf{x}, A\mathbf{x}) - \tfrac{1}{2}(\mathbf{b}, \mathbf{x}^*) + c_0 \\ &= \tfrac{1}{2}(\mathbf{x}, A\mathbf{x}) - (\mathbf{x}, \mathbf{x}) + c_0 \equiv F(\mathbf{x}) \end{aligned}$$

Thus, the minimizer of $f(\mathbf{x})$ and that of $F(\mathbf{x})$ coincide, provided that $\mathbf{x}^*$ is the exact solution of $A\mathbf{x} = \mathbf{b}$.

Alternative view-point, cont.

$$F(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x}, A\mathbf{x}) - (\mathbf{x}, \mathbf{x}) + c_0$$

We decide to compute the minimization problem for $F(\mathbf{x})$ and to do it iteratively, locally per iteration, performing a line search, namely,
we seek $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \tau_k \mathbf{d}^k$ such that F will be minimized.

How to choose τ_k and $\mathbf{d}^k$?

Alternative view-point, cont.

Theorem 1:
Let $F(\mathbf{x}) \in C^1)R^n)$ and let ∇F be the gradient of F at some point $\mathbf{x}$.
If $(\nabla F, \mathbf{d}) < 0$, then $\mathbf{d}$ is a descent direction for F at $\mathbf{x}$.
Proof: Descent direction: $F(\mathbf{x} + \tau\mathbf{d}) \leq F(\mathbf{x})$ for $0 \leq \tau \leq \tau_0$

$$F(\mathbf{x} + \tau\mathbf{d}) = F(\mathbf{x}) + \tau \underbrace{(\nabla F, \mathbf{d})}_{<0} + O(\tau)$$

Thus, τ can be chosen small enough so that
$\tau(\nabla F, \mathbf{d}) + O(\tau) < 0$

Alternative view-point, cont.

Theorem 2:
Among all search directions $\mathbf{d}$ at some point $\mathbf{x}$, F descents most rapidly for $\mathbf{d} = \nabla F$.
Proof: We want to minimize the directional derivative of F at $\mathbf{x}$ over all possible search directions.
The (first) directional derivative in direction $\mathbf{y}$ at $\mathbf{x}$ is defined as follows:

$$\frac{dF}{d\mathbf{y}} = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i} y_i = (\nabla F, \mathbf{y}).$$

Let $\mathbf{y}$ be arbitrary, $|\mathbf{y}\| = 1$.

$$|(\nabla F, \mathbf{y})| \leq \|\nabla F\| \|\mathbf{y}\| = \|\nabla F\|$$

Thus, there holds $|(\nabla F, \mathbf{y})| \geq -\|\nabla F\|$.
For the special choice $\mathbf{y} = -\nabla F / \|\nabla F\|$ we obtain
$(\nabla F, -\nabla F / \|\nabla F\|) = -\|\nabla F\|$.