Conjugate Gradient and Preconditioned Conjugate Gradient iterations

Let $\mathbf{e}_k$ be the error at step k of an iterative method in approximating the solution $\mathbf{x}$ of a linear system $A\mathbf{x} = \mathbf{b}$, i.e. $\mathbf{e}_k = \mathbf{x} - \mathbf{x}_k$. Choose a vector $\mathbf{d}_k \neq \mathbf{0}$ and set $\mathbf{e}_{k+1} = \mathbf{e}_k - \omega \mathbf{d}_k$. Let H be a positive definite matrix and consider the inner product $(\mathbf{u}, \mathbf{v})_H = \mathbf{u}^T H \mathbf{v}$. Then the value of ω for which $\|\mathbf{e}_{k+1}\|_H$ is minimum is

$$\omega = \omega_k = \frac{(\mathbf{e}_k, \mathbf{d}_k)_H}{\|\mathbf{d}_k\|_H^2}.$$

($\|\mathbf{e}_{k+1}\|_H^2 = \|\mathbf{e}_k\|^2 - 2\omega(\mathbf{e}_k, \mathbf{d}_k)_H + \omega^2 \|\mathbf{d}_k\|_H^2$). So, we have the iterative scheme

$$\mathbf{x}_0 \in \mathbb{R}^n, \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \frac{(\mathbf{e}_k, \mathbf{d}_k)_H}{\|\mathbf{d}_k\|_H^2} \mathbf{d}_k, \quad k = 0, 1, 2, \ldots . \tag{it}$$

Note that each of the three conditions:

1) $\|\mathbf{e}_{k+1}\|_H$ minimum,
2) $(\mathbf{e}_{k+1}, \mathbf{d}_k)_H = 0$,
3) $F(\mathbf{x}_k + \omega \mathbf{d}_k)$ minimum, $F(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T H \mathbf{y} - \mathbf{y}^T H A^{-1} \mathbf{b}$

yields the value $\omega = \omega_k$, i.e. such conditions are equivalent. (Note that $A^{-1}\mathbf{b}$ is the global minimum for F, and the contours of F are neighborhoods of $A^{-1}\mathbf{b}$ in the metric induced by the norm $\|\cdot\|_H$). Moreover, for $\omega = \omega_k$ we have

4) $\|\mathbf{e}_{k+1}\|_H^2 = \|\mathbf{e}_k\|_H^2 - \|\omega_k \mathbf{d}_k\|_H^2$,
5) $\lim_k \|\mathbf{e}_k\|_H = l_{\{\mathbf{d}_k\},H} \geq 0$,
6) $\lim_{k \to +\infty} \|\omega_k \mathbf{d}_k\|_H = 0$.

Suitable choice of H and $\{\mathbf{d}_k\}$ make $l_{\{\mathbf{d}_k\},H} = 0$, i.e. make (it) convergent to $\mathbf{x} = A^{-1}\mathbf{b}$.

Choice $H = A^T A$ ... $H = I$ ... ...

Choice $H = A$.

G method

Assume A, in the system $A\mathbf{x} = \mathbf{b}$ we have to solve, positive definite. Then the choices $H = A$ and $\mathbf{d}_k = \mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$ yield $l_{\{\mathbf{d}_k\},H} = 0$ (use 6)). The method so obtained is called steepest descent (or Gradient) since $\mathbf{d}_k = \mathbf{r}_k = -\nabla F(\mathbf{x}_k)$ and F decreases along $-\nabla F(\mathbf{x}_k)$ more rapidly than in any other direction (in a neighborhood of $\mathbf{x}_k$). However, in general the contours of F are far from being spheres, so the steepest descent direction (which is orthogonal to such contours) is far from pointing to $A^{-1}\mathbf{b}$. In particular, from 4) and the Kantorovich inequality,

$$1 \leq \frac{\mathbf{z}^T A \mathbf{z} \mathbf{z}^T A^{-1} \mathbf{z}}{(\mathbf{z}^T \mathbf{z})^2} \leq \frac{(\lambda_{\max} + \lambda_{\min})^2}{4 \lambda_{\max} \lambda_{\min}}$$

($\lambda_{\max} = \max \lambda(A)$, $\lambda_{\min} = \min \lambda(A)$), we have the following result

$$\frac{\|\mathbf{x} - \mathbf{x}_k\|_A}{\|\mathbf{x} - \mathbf{x}_0\|_A} \leq \left(\frac{\frac{\lambda_{\max}}{\lambda_{\min}} - 1}{\frac{\lambda_{\max}}{\lambda_{\min}} + 1} \right)^k$$

that states that G can be very slow when $\frac{\lambda_{\max}}{\lambda_{\min}} >> 1$.

CG method

Assume A, the coefficient matrix of our system $A\mathbf{x} = \mathbf{b}$, positive definite (recall that A and $\mathbf{b}$ are real). In the general scheme choose $H = A$, $\mathbf{d}_0 = \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$, $\mathbf{d}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{d}_{k-1}$, $k = 1, 2, \ldots$, $(\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k)$ where β_{k-1} is such that

$$(\mathbf{d}_k, \mathbf{d}_{k-1})_A = 0$$

($\mathbf{d}_k$ *conjugate to* $\mathbf{d}_{k-1}$). Here below is the algorithm we obtain:

$$\begin{array}{l}
\mathbf{x}_0 \in \mathbb{R}^n,\ \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0,\ \mathbf{d}_0 = \mathbf{r}_0. \\
\text{For } k = 0, 1, \ldots, \{ \\
\tau_k = \frac{\mathbf{d}_k^T \mathbf{r}_k}{\mathbf{d}_k^T A \mathbf{d}_k} \\
\mathbf{x}_{k+1} = \mathbf{x}_k + \tau_k \mathbf{d}_k \\
\mathbf{r}_{k+1} = \mathbf{b} - A\mathbf{x}_{k+1} = \mathbf{r}_k - \tau_k A\mathbf{d}_k \\
\beta_k = -\frac{\mathbf{r}_{k+1}^T A \mathbf{d}_k}{\mathbf{d}_k^T A \mathbf{d}_k} \\
\mathbf{d}_{k+1} = \mathbf{r}_{k+1} + \beta_k \mathbf{d}_k \\
\}
\end{array}$$

known as Conjugate Gradient (CG) algorithm.

Remarks. Note that

$$0 = (\mathbf{x} - \mathbf{x}_{k+1})^T H \mathbf{d}_k = (\mathbf{x} - \mathbf{x}_{k+1})^T A \mathbf{d}_k = \mathbf{r}_{k+1}^T \mathbf{d}_k,\ \mathbf{r}_{k+1}^T \mathbf{d}_{k+1} = \|\mathbf{r}_{k+1}\|_2^2.$$

As a consequence, if at step s we have $\mathbf{r}_s = \mathbf{b} - A\mathbf{x}_s \neq \mathbf{0}$, then $\mathbf{d}_s \neq \mathbf{0}$, τ_s is well defined and not zero ...: the algorithm works.

If $\mathbf{r}_0, \ldots, \mathbf{r}_{m-1}$ are non null and $\mathbf{r}_m = \mathbf{0}$, then $\beta_{m-1} = 0$, $\mathbf{d}_m = \mathbf{0}$, τ_m cannot be defined, but it doen't matter since $\mathbf{x}_m = A^{-1}\mathbf{b}$. This hypothesis is effectively verified, in fact there exists $m \leq n =$ the order of A, such that $\mathbf{r}_m = \mathbf{0}$ (see below).

Alternative expressions for τ_k and β_k hold:

$$\tau_k = \frac{\mathbf{d}_k^T \mathbf{r}_k}{\mathbf{d}_k^T A \mathbf{d}_k} = \frac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{d}_k^T A \mathbf{d}_k}$$

($\mathbf{d}_k = \mathbf{r}_k + \beta_{k-1}\mathbf{d}_{k-1}$, $\mathbf{r}_k^T \mathbf{d}_{k-1} = 0$),

$$\begin{array}{rcl}
\beta_k & = & -\frac{\mathbf{r}_{k+1}^T A \mathbf{d}_k}{\mathbf{d}_k^T A \mathbf{d}_k} = -\frac{\mathbf{r}_{k+1}^T \tau_k^{-1}(\mathbf{r}_k - \mathbf{r}_{k+1})}{\mathbf{d}_k^T \tau_k^{-1}(\mathbf{r}_k - \mathbf{r}_{k+1})} \\
& = & \frac{-\mathbf{r}_{k+1}^T \mathbf{r}_k + \mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{d}_k^T \mathbf{r}_k} = \frac{\mathbf{r}_{k+1}^T \mathbf{r}_{k+1}}{\mathbf{r}_k^T \mathbf{r}_k}.
\end{array}$$

The latter identity uses the result

$$\mathbf{r}_{k+1}^T \mathbf{r}_k = 0$$

(residual at step $k+1$ is orthogonal to residual at step k, exactly as in the Gradient method) which is not obvious:

$$\begin{array}{rcl}
\mathbf{r}_{k+1}^T \mathbf{r}_k & = & \mathbf{r}_{k+1}^T (\mathbf{d}_k - \beta_{k-1}\mathbf{d}_{k-1}) = -\beta_{k-1} \mathbf{r}_{k+1}^T \mathbf{d}_{k-1} \\
& = & -\beta_{k-1}(\mathbf{r}_k - \tau_k A\mathbf{d}_k)^T \mathbf{d}_{k-1} = \beta_{k-1}\tau_k \mathbf{d}_k^T A \mathbf{d}_{k-1} = 0.
\end{array}$$

First main result: If $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_p$ are non null, then

$$\mathbf{d}_l^T A \mathbf{d}_j = 0,\ \ \mathbf{r}_l^T \mathbf{r}_j = 0,\ \ 0 \leq j < l \leq p.$$

That is, each new residual (search direction) is orthogonal (conjugate) to all previous residuals (search directions). As a consequence, the residual $\mathbf{r}_m$ must be null for some $m \leq n$, or, equivalently, CG finds the solution of $A\mathbf{x} = \mathbf{b}$ in at most n steps.

Proof. ...

A useful representation of the residuals. If $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_{k-1}$ are non null, then there exist polynomials $s_k(\lambda), q_k(\lambda)$ such that

$$\begin{gathered} \mathbf{r}_k = s_k(A)\mathbf{r}_0, \quad \mathbf{d}_k = q_k(A)\mathbf{r}_0, \\ s_k(\lambda) = (-1)^k \tau_0\tau_1\cdots\tau_{k-1}\lambda^k + \ldots + 1, \quad \tau_0\tau_1\cdots\tau_{k-1} \neq 0. \end{gathered}$$

Proof (by induction). The equality $\mathbf{r}_0 = s_0(A)\mathbf{r}_0$ holds if $s_0(\lambda) = 1$; $\mathbf{d}_0 = q_0(A)\mathbf{r}_0$ holds if $q_0(\lambda) = 1$. Moreover,

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \tau_k A\mathbf{d}_k = s_k(A)\mathbf{r}_0 - \tau_k A q_k(A)\mathbf{r}_0 = s_{k+1}(A)\mathbf{r}_0$$

if $s_{k+1}(\lambda) = s_k(\lambda) - \tau_k\lambda q_k(\lambda)$, and

$$\mathbf{d}_{k+1} = \mathbf{r}_{k+1} + \beta_k\mathbf{d}_k = s_{k+1}(A)\mathbf{r}_0 + \beta_k q_k(A)\mathbf{r}_0 = q_{k+1}(A)\mathbf{r}_0$$

if $q_{k+1}(\lambda) = s_{k+1}(\lambda) + \beta_k q_k(\lambda)$. Finally, since

$$s_{k+1}(\lambda) = s_k(\lambda) - \tau_k\lambda(s_k(\lambda) + \beta_{k-1}q_{k-1}(\lambda)),$$

the coefficient of λ^{k+1} in $s_{k+1}(\lambda)$ is $-\tau_k$ times the coefficient of λ^k in $s_k(\lambda)$. Thus, by the inductive assumption, it must be $(-1)^{k+1}\tau_0\tau_1\cdots\tau_{k-1}\tau_k$. Also, the coefficient of λ^0 in $s_{k+1}(\lambda)$ is equal to the coefficient of λ^0 in $s_k(\lambda)$, which is 1 by the inductive assumption.

Second main result: $\mathbf{r}_k = \mathbf{0}$ for some $k \leq \#\{\text{distinct eigenvalues of } A\}$.

Proof. Let $\mu_1, \mu_2, \ldots, \mu_m$ be the distinct eigenvalues of A ($m \leq n =$ order of A). Assume that CG requires more than m steps to converge. So, the vectors $\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_m$ are non null, and, by the First main result, orthogonal ($\Rightarrow$ linearly independent). Let V be an orthonormal matrix whose columns are eigenvectors of A, thus $V^T = V^{-1}$ and $AV = VD$ for D diagonal with the eigenvalues of A as diagonal entries. Observe that there is a degree-m polynomial which is null in A,

$$\prod_{j=1}^{m}(A-\mu_j I) = \prod_{j=1}^{m}(VDV^T-\mu_j I) = \prod_{j=1}^{m} V(D-\mu_j I)V^T == V\prod_{j=1}^{m}(D-\mu_j I)V^T = 0.$$

As a consequence the matrices $A^0 = I$, A, ..., A^m are linearly dependent. But this implies that the dimension of the space

$$S_{m+1}(\mathbf{r}_0) = \text{Span}\,\{\mathbf{r}_0, A\mathbf{r}_0, \ldots, A^m\mathbf{r}_0\} = \text{Span}\,\{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_m\}$$

is smaller than $m + 1$, which is absurd. It follows that one of the vectors $\mathbf{r}_i$, $i = 0, \ldots, m$, must be null.

Let Π_k^1 be the set of all polynomials of degree exactly k whose graphic pass through $(0, 1)$. We now see that the polynomial $s_k(\lambda)$ in the expression $\mathbf{r}_k = s_k(A)\mathbf{r}_0$ is a very particular polynomial in the class Π_k^1: it makes the norm

of the vector $p_k(A)\mathbf{r}_0$, $p_k \in \Pi_k^1$, minimum (for a suitable choice of the norm). This result let us give estimates of the rate of convergence of CG, as precise as good is the knowledge about the location of the eigenvalues of A. For example, if it is known that the eigenvalues of A *cluster* around 1, then CG must converge with a superlinear rate of convergence (see toe_1a).

Notice that $\mathbf{r}_k = s_k(A)\mathbf{r}_0 = \mathbf{r}_0 + \hat{\mathbf{h}}_k$, for a particular vector $\hat{\mathbf{h}}_k$ in the space $\mathcal{M} = \text{Span}\{A\mathbf{r}_0, A^2\mathbf{r}_0, \ldots, A^k\mathbf{r}_0\}$. Take a generic vector $\mathbf{h}_k$ in this space. Then

$$\begin{aligned} \|\mathbf{r}_0 + \mathbf{h}_k\|_{A^{-1}}^2 &= \|\mathbf{r}_0 + \hat{\mathbf{h}}_k + \mathbf{h}_k - \hat{\mathbf{h}}_k\|_{A^{-1}}^2 \\ &= \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|_{A^{-1}}^2 + \|\mathbf{h}_k - \hat{\mathbf{h}}_k\|_{A^{-1}}^2 + 2(\mathbf{r}_0 + \hat{\mathbf{h}}_k, \mathbf{h}_k - \hat{\mathbf{h}}_k)_{A^{-1}}. \end{aligned}$$

Now observe that the latter inner product is null, in fact, for $j = 0, \ldots, k-1$, $0 = \mathbf{r}_k^T\mathbf{r}_j = \mathbf{r}_k^T A^{-1} A\mathbf{r}_j = (\mathbf{r}_k, A\mathbf{r}_j)_{A^{-1}}$, that is, $\mathbf{r}_k$ is A^{-1}-orthogonal to the space $\text{Span}\{A\mathbf{r}_0, A\mathbf{r}_1, \ldots, A\mathbf{r}_{k-1}\}$, but this space is exactly $\mathcal{M}$. The thesis follows since $\mathbf{h}_k - \hat{\mathbf{h}}_k \in \mathcal{M}$. So we have:

$$\|\mathbf{r}_0 + \mathbf{h}_k\|_{A^{-1}}^2 = \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|_{A^{-1}}^2 + \|\mathbf{h}_k - \hat{\mathbf{h}}_k\|_{A^{-1}}^2 \geq \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|_{A^{-1}}^2.$$

In other words,

$$\begin{aligned} \|\mathbf{r}_k\|_{A^{-1}}^2 &= \|\mathbf{r}_0 + \hat{\mathbf{h}}_k\|_{A^{-1}}^2 = \min\{\|\mathbf{r}_0 + \mathbf{h}_k\|_{A^{-1}}^2 : \mathbf{h}_k \in \mathcal{M}\} \\ &= \min\{\|p_k(A)\mathbf{r}_0\|_{A^{-1}}^2 : p_k \in \Pi_k^1\}. \end{aligned} \qquad (m)$$

Comparison with GMRES. Notice that for any $\mathbf{h}_k \in \mathcal{M}$ we have

$$\mathbf{r}_0 + \mathbf{h}_k = \mathbf{b} - A(\mathbf{x}_0 + \mathbf{z}),\ \mathbf{z} = -A^{-1}\mathbf{h}_k \in \mathcal{S}_k(\mathbf{r}_0) = \text{Span}\{\mathbf{r}_0, A\mathbf{r}_0, \ldots, A^{k-1}\mathbf{r}_0\}.$$

Thus, the vector $\mathbf{x}_k$ generated by the CG method is of type $\mathbf{x}_0 + \hat{\mathbf{z}}$ where $\hat{\mathbf{z}}$ solves the problem

$$\|\mathbf{b} - A(\mathbf{x}_0 + \hat{\mathbf{z}})\|_{A^{-1}} = \min\{\|\mathbf{b} - A(\mathbf{x}_0 + \mathbf{z})\|_{A^{-1}} : \mathbf{z} \in \mathcal{S}_k(\mathbf{r}_0)\} \qquad (p)$$

($\mathcal{S}_k(\mathbf{r}_0)$ is known as Krilov space). GMRES is a method able to solve $A\mathbf{x} = \mathbf{b}$ in at most n steps under the only assumption $\det(A) \neq 0$. (Like CG, GMRES in order to be competitive must be used as an iterative method, i.e. less than n steps must be sufficient to give a good approximation of $\mathbf{x}$). In the k-th step of GMRES it is defined a vector $\mathbf{x}_k$ of type $\mathbf{x}_0 + \hat{\mathbf{z}}$ where $\hat{\mathbf{z}}$ solves exactly the problem (p) but the norm involved is the euclidean one. So, CG is a minimal residual algorithm different from GMRES$|_{A\,pd}$.

It is easy to see that the condition (m) can be rewritten as follows:

$$\|\mathbf{x} - \mathbf{x}_k\|_A^2 = \min_{p_k \in \Pi_k^1} \|p_k(A)(\mathbf{x} - \mathbf{x}_0)\|_A^2.$$

Now we give a bound for the quantity $\|p_k(A)(\mathbf{x} - \mathbf{x}_0)\|_A^2$, $p_k \in \Pi_k^1$, which can be evaluated if (besides $A, \mathbf{b}$) also some information about the location of the eigenvalues λ_i of A is given. Let $\mathbf{v}_i \neq \mathbf{0}$ be such that $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$, $\mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$. Then

$$\begin{aligned} \|p_k(A)(\mathbf{x} - \mathbf{x}_0)\|_A^2 &= (\mathbf{x} - \mathbf{x}_0)^T A p_k(A)^2(\mathbf{x} - \mathbf{x}_0) = (\textstyle\sum \alpha_i\mathbf{v}_i)^T \sum \alpha_i A p_k(A)^2\mathbf{v}_i \\ &= (\textstyle\sum \alpha_i\mathbf{v}_i)^T \sum \alpha_i A p_k(\lambda_i)^2\mathbf{v}_i = (\textstyle\sum \alpha_i\mathbf{v}_i)^T \sum \alpha_i\lambda_i p_k(\lambda_i)^2\mathbf{v}_i \\ &= \textstyle\sum \alpha_i^2\lambda_i p_k(\lambda_i)^2 \leq \max_i |p_k(\lambda_i)|^2 \|\mathbf{x} - \mathbf{x}_0\|_A^2. \end{aligned}$$

So, we obtain the following

Third main result: If $\mathbf{x}_k$ is the k-th vector generated by CG when applied to solve the pd linear system $A\mathbf{x} = \mathbf{b}$, then

$$\|\mathbf{x} - \mathbf{x}_k\|_A^2 = \min_{p_k \in \Pi_k^1} \|p_k(A)(\mathbf{x} - \mathbf{x}_0)\|_A^2 \leq \max_i |p_k(\lambda_i)|^2 \|\mathbf{x} - \mathbf{x}_0\|_A^2, \quad \forall\, p_k \in \Pi_k^1.$$

So, if $S \subset \mathbb{R}$, $p_k \in \Pi_k^1$, $M_k \in \mathbb{R}$ are known such that $\lambda_i \in S\ \forall\, i$ and $|p_k(\lambda)| \leq M_k$ $\forall\, \lambda \in S$, then $\|\mathbf{x} - \mathbf{x}_k\|_A \leq M_k \|\mathbf{x} - \mathbf{x}_0\|_A$.

Let us see two applications of the latter result. As consequences of the first application we observe that CG (considered as an iterative method) has a linear rate of convergence, is in general faster than G, and is competitive (f.i. with direct methods) if $\lambda_{\max}$ and $\lambda_{\min}$ are comparable. However, as a consequence of the second application, the latter condition is not necessary: the rate of convergence of CG remains high (so, CG remains competitive) if most of the eigenvalues are in $[\lambda_{\min}, \hat{\lambda}]$ with $\lambda_{\min}$ and $\hat{\lambda}$ comparable. Further useful applications of the Third main result hold. In particular, as a consequence of one of these (see toe_1a), it can be stated that CG has a superlinear rate of convergence if most of the eigenvalues of A are in the interval $S = [1 - \varepsilon, 1 + \varepsilon]$ $(\ldots)$.

(1)

$$S = [\lambda_{\min}, \lambda_{\max}],\ p_k(x) = \frac{T_k\left(\frac{\lambda_{\max} + \lambda_{\min} - 2x}{\lambda_{\max} - \lambda_{\min}}\right)}{T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)} \ \Rightarrow$$

$$\|\mathbf{x} - \mathbf{x}_k\|_A < 2\left(\frac{\sqrt{\mu_2(A)} - 1}{\sqrt{\mu_2(A)} + 1}\right)^k \|\mathbf{x} - \mathbf{x}_0\|_A,\ \mu_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

(2)

$$S = [\lambda_{\min}, \hat{\lambda}] \cup \{\lambda_i :\ \lambda_i > \hat{\lambda}\},\ r_{\hat{\lambda}} = \#\{i :\ \lambda_i > \hat{\lambda}\},$$

$$p_k(x) = \Pi_{i:\,\lambda_i > \hat{\lambda}}\left(1 - \frac{x}{\lambda_i}\right) \frac{T_{k - r_{\hat{\lambda}}}\left(\frac{\hat{\lambda} + \lambda_{\min} - 2x}{\hat{\lambda} - \lambda_{\min}}\right)}{T_{k - r_{\hat{\lambda}}}\left(\frac{\hat{\lambda} + \lambda_{\min}}{\hat{\lambda} - \lambda_{\min}}\right)} \ \Rightarrow$$

$$\|\mathbf{x} - \mathbf{x}_k\|_A < 2\left(\frac{\sqrt{\hat{\lambda}/\lambda_{\min}} - 1}{\sqrt{\hat{\lambda}/\lambda_{\min}} + 1}\right)^{k - r_{\hat{\lambda}}} \|\mathbf{x} - \mathbf{x}_0\|_A, \quad k \geq r_{\hat{\lambda}}.$$

The applications (1) and (2) of the Third main result suggest an idea. When $\lambda_{\min}$ and $\lambda_{\max}$ are not comparable and the eigenvalues of A are uniformly distributed in the interval $[\lambda_{\min}, \lambda_{\max}]$ (in this case all n steps of CG are required in order to give a good approximation of $\mathbf{x}$), replace the given system $A\mathbf{x} = \mathbf{b}$ with an equivalent system $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$, $\tilde{A} = E^{-1}AE^{-T}$, $\tilde{\mathbf{x}} = E^T\mathbf{x}$, $\tilde{\mathbf{b}} = E^{-1}\mathbf{b}$, $\det(E) \neq 0$, where the matrix E is such that $\mu_2(\tilde{A}) < \mu_2(A)$ and has one of the following properties

- $\mu_2(\tilde{A}) << \mu_2(A)$
- $\tilde{A}$ has much less distinct eigenvalues than A

- $\tilde{A}$ has the eigenvalues much more clustered (around 1) than A

Then apply CG to $\tilde{A}\tilde{x} = \tilde{\mathbf{b}}$.

If such matrix E can be found, then the pd matrix $P = EE^T$ is said *preconditioner*.

Note that $E^{-T}\tilde{A}E^T = P^{-1}A$, so one could look directly for a pd matrix P such that the (real positive) eigenvalues of $P^{-1}A$ have the required properties. For example, in order to obtain something of type $P^{-1}A \approx I$ (which would result in a very high increase of the CG rate of convergence) one could choose P as an approximation $\mathcal{A}$ of A. We shall see that applying CG to $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$ requires, for each step, a surplus of computation: solve a system of type $P\mathbf{z} = \mathbf{h}_k$. This computation must not make CG slow, in other words P must be a lower complexiy matrix than A. Also notice that E_1 and E_2, $E_1 \neq E_2$, $E_1E_1^T = E_2E_2^T$, define matrices $\tilde{A}_1 = E_1^{-1}AE_1^{-T}$ and $\tilde{A}_2 = E_2^{-1}AE_2^{-T}$, $\tilde{A}_1 \neq \tilde{A}_2$, with the same spectrum. For this reason one prefers to call preconditioner P instead of E.

A final remark. The vector $\mathbf{x} = A^{-1}\mathbf{b}$ we are looking for is also the minimum point of the function $F(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T A\mathbf{z} - \mathbf{z}^T\mathbf{b}$. Analogously, $\tilde{\mathbf{x}} = \tilde{A}^{-1}\tilde{\mathbf{b}}$ is the minimum point of the function $\tilde{F}(\mathbf{z}) = \frac{1}{2}\mathbf{z}^T\tilde{A}\mathbf{z} - \mathbf{z}^T\tilde{\mathbf{b}}$. The preconditioning technique replaces the (sections of the) contours of F with the more spherical (sections of the) contours of $\tilde{F}$, and this results in a more efficient minimization when using gradient-type methods.

Let us write the preconditioned version of the CG algorithm, well defined once that A, $\mathbf{b}$ and the preconditioner P are given.

Let us apply CG to the system $\tilde{A}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$:

$$
\begin{array}{l}
\tilde{\mathbf{x}}_0 \in \mathbb{R}^n,\ \tilde{\mathbf{r}}_0 = \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}_0,\ \tilde{\mathbf{d}}_0 = \tilde{\mathbf{r}}_0. \\
\text{For } k = 0, 1, \ldots, \{ \\
\tilde{\tau}_k = \frac{\tilde{\mathbf{r}}_k^T\tilde{\mathbf{r}}_k}{\tilde{\mathbf{d}}_k^T\tilde{A}\tilde{\mathbf{d}}_k} \\
\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{x}}_k + \tilde{\tau}_k\tilde{\mathbf{d}}_k \\
\tilde{\mathbf{r}}_{k+1} = \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{r}}_k - \tilde{\tau}_k\tilde{A}\tilde{\mathbf{d}}_k \\
\tilde{\beta}_k = \frac{\tilde{\mathbf{r}}_{k+1}^T\tilde{\mathbf{r}}_{k+1}}{\tilde{\mathbf{r}}_k^T\tilde{\mathbf{r}}_k} \\
\tilde{\mathbf{d}}_{k+1} = \tilde{\mathbf{r}}_{k+1} + \tilde{\beta}_k\tilde{\mathbf{d}}_k \\
\}
\end{array}
$$

Note that the convergence rate of the sequence $\{\tilde{\mathbf{x}}_k\}$ can be evaluated by using the following results

$$
\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_k\|_{\tilde{A}} < 2\left(\frac{\sqrt{\mu_2(\tilde{A})} - 1}{\sqrt{\mu_2(\tilde{A})} + 1}\right)^k \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_0\|_{\tilde{A}}, \quad \mu_2(\tilde{A}) = \frac{\tilde{\lambda}_{\max}}{\tilde{\lambda}_{\min}},
$$

$$
\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_k\|_{\tilde{A}} < 2\left(\frac{\sqrt{\tilde{\tilde{\lambda}}/\tilde{\lambda}_{\min}} - 1}{\sqrt{\tilde{\tilde{\lambda}}/\tilde{\lambda}_{\min}} + 1}\right)^{k - r_{\tilde{\tilde{\lambda}}}} \|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_0\|_{\tilde{A}}, \quad k \geq r_{\tilde{\tilde{\lambda}}} :
$$

if $\mu_2(\tilde{A}) << \mu_2(A)$ or $\tilde{A}$ has most of the eigenvalues $\tilde{\lambda}_i$ in $[\tilde{\lambda}_{\min}, \tilde{\tilde{\lambda}}]$ and $\tilde{\tilde{\lambda}}/\tilde{\lambda}_{\min} << \lambda_{\max}/\lambda_{\min}$, then $\tilde{\mathbf{x}}_k \to \tilde{\mathbf{x}} = E^T\mathbf{x}$ with a greater rate than $\mathbf{x}_k \to \mathbf{x}$.

Now we obtain each row of the preconditioned CG method. Define $\mathbf{x}_k = E^{-T}\tilde{\mathbf{x}}_k$, $\mathbf{r}_k = \mathbf{b} - A\mathbf{x}_k$, and $\mathbf{d}_k = E^{-T}\tilde{\mathbf{d}}_k$. Then

$$\begin{aligned}\tilde{\mathbf{r}}_k &= \tilde{\mathbf{b}} - \tilde{A}\tilde{\mathbf{x}}_k = E^{-1}\mathbf{b} - E^{-1}AE^{-T}(E^T\mathbf{x}_k)\\ &= E^{-1}\mathbf{r}_k = E^TE^{-T}E^{-1}\mathbf{r}_k = E^T\mathbf{h}_k, \mathbf{h}_k = P^{-1}\mathbf{r}_k,\end{aligned}$$

$$\begin{aligned}\tilde{\mathbf{r}}_k^T\tilde{\mathbf{r}}_k &= \mathbf{r}_k^TE^{-T}E^{-1}\mathbf{r}_k = \mathbf{r}_k^T\mathbf{h}_k,\\ \tilde{\mathbf{d}}_k^T\tilde{A}\tilde{\mathbf{d}}_k &= \tilde{\mathbf{d}}_k^TE^{-1}AE^{-T}\tilde{\mathbf{d}}_k = \mathbf{d}_k^TA\mathbf{d}_k.\end{aligned}$$

Thus

$$\tilde{\tau}_k = \frac{\mathbf{r}_k^T\mathbf{h}_k}{\mathbf{d}_k^TA\mathbf{d}_k}. \tag{row1}$$

Moreover, we have

$$\begin{aligned}&\tilde{\mathbf{x}}_{k+1} = E^T\mathbf{x}_{k+1} = E^T\mathbf{x}_k + \tilde{\tau}_kE^T\mathbf{d}_k \Rightarrow\\ &\mathbf{x}_{k+1} = \mathbf{x}_k + \tilde{\tau}_k\mathbf{d}_k,\end{aligned} \tag{row2}$$

$$\begin{aligned}&\tilde{\mathbf{r}}_{k+1} = E^{-1}\mathbf{r}_{k+1} = E^{-1}\mathbf{r}_k - \tilde{\tau}_kE^{-1}AE^{-T}E^T\mathbf{d}_k \Rightarrow\\ &\mathbf{r}_{k+1} = \mathbf{r}_k + \tilde{\tau}_kA\mathbf{d}_k,\end{aligned} \tag{row3}$$

$$\tilde{\beta}_k = \frac{\mathbf{r}_{k+1}^T\mathbf{h}_{k+1}}{\mathbf{r}_k^T\mathbf{h}_k} \tag{row4}$$

(row3.5: $\mathbf{h}_{k+1} = P^{-1}\mathbf{r}_{k+1}$),

$$\begin{aligned}&\tilde{\mathbf{d}}_{k+1} = E^T\mathbf{d}_{k+1} = E^T\mathbf{h}_{k+1} + \tilde{\beta}_kE^T\mathbf{d}_k \Rightarrow\\ &\mathbf{d}_{k+1} = \mathbf{h}_{k+1} + \tilde{\beta}_k\mathbf{d}_k.\end{aligned} \tag{row5}$$

Finally, in order to initialize the algorithm, set:

$$\begin{aligned}&\mathbf{x}_0 = E^{-T}\tilde{\mathbf{x}}_0,\ \mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0,\\ &\mathbf{d}_0 = E^{-T}\tilde{\mathbf{d}}_0 = E^{-T}\tilde{\mathbf{r}}_0 = E^{-T}E^T\mathbf{h}_0 = \mathbf{h}_0.\end{aligned} \tag{row0}$$

Regarding the convergence rate of the sequence $\{\mathbf{x}_k\}$, generated by the algorithm row0 and, for $k = 0, 1, \ldots$, rows$1, 2, 3, 3.5, 4, 5$, note that

$$\begin{aligned}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}\|_{\tilde{A}}^2 &= (\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}})^T\tilde{A}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}) = (E^T\mathbf{x}_k - E^T\mathbf{x})^TE^{-1}AE^{-T}(E^T\mathbf{x}_k - E^T\mathbf{x})\\ &= (\mathbf{x}_k - \mathbf{x})^TA(\mathbf{x}_k - \mathbf{x}) = \|\mathbf{x}_k - \mathbf{x}\|_A^2.\end{aligned}$$

Thus the bounds for $\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_k\|_{\tilde{A}}$ obtained above, can be rewritten as follows

$$\begin{aligned}&\frac{\|\mathbf{x}_k - \mathbf{x}\|_A}{\|\mathbf{x}_0 - \mathbf{x}\|_A} \le 2\left(\frac{\sqrt{\mu_2(\tilde{A})}-1}{\sqrt{\mu_2(\tilde{A})}+1}\right)^k, \quad \mu_2(\tilde{A}) = \frac{\tilde{\lambda}_{\max}}{\tilde{\lambda}_{\min}},\\ &\frac{\|\mathbf{x}_k - \mathbf{x}\|_A}{\|\mathbf{x}_0 - \mathbf{x}\|_A} \le 2\left(\frac{\sqrt{\tilde{\lambda}/\tilde{\lambda}_{\min}}-1}{\sqrt{\tilde{\lambda}/\tilde{\lambda}_{\min}}+1}\right)^{k-r_{\tilde{\lambda}}}, \quad k \ge r_{\tilde{\lambda}}.\end{aligned}$$

Why clustering around 1 *is good*

Let A be a p.d. matrix and ε, $0 < \varepsilon < 1$, be fixed.

Denote by λ_j^ε the eigenvalues of A outside the interval $[1-\varepsilon, 1+\varepsilon]$ and by r_ε the number of such eigenvalues. Set $S = [1-\varepsilon, 1+\varepsilon] \cup \{\lambda_j^\varepsilon\}$ and let p_q be the polynomial

$$p_q(\lambda) = \prod_{\lambda_j^\varepsilon}\left(1 - \frac{\lambda}{\lambda_j^\varepsilon}\right)\frac{T_{q-r_\varepsilon}((1-\lambda)/\varepsilon)}{T_{q-r_\varepsilon}(1/\varepsilon)},\ q \ge r_\varepsilon$$

where $T_k(x)$ denotes the chebycev polynomial of degree k. $((b+a-2\lambda)/(b-a) = (1-\lambda)/\varepsilon$, $(b+a)/(b-a) = 1/\varepsilon$, if $a = 1-\varepsilon$, $b = 1+\varepsilon$). Notice that S is a set containing all the eigenvalues of A, and p_q has exactly degree q and $p_q(0) = 1$. Then one can say that if $\mathbf{x}_q$ is the q-th vector generated by the CG method when solving $A\mathbf{x} = \mathbf{b}$, then

$$\|\mathbf{x} - \mathbf{x}_q\|_A \leq (\max_{\lambda \in S} |p_q(\lambda)|)\|\mathbf{x} - \mathbf{x}_0\|_A. \qquad (bound)$$

This bound for $\|\mathbf{x} - \mathbf{x}_q\|_A$ allows a better evaluation of the CG rate of convergence with respect to the well known bound

$$\|\mathbf{x} - \mathbf{x}_q\|_A \leq 2\left(\frac{\sqrt{\mu_2(A)} - 1}{\sqrt{\mu_2(A)} + 1}\right)^q \|\mathbf{x} - \mathbf{x}_0\|_A, \quad \mu_2(A) = \frac{\max \lambda(A)}{\min \lambda(A)} \quad (wkbound)$$

in case it is known that most of (almost all) the eigenvalues of A are in some interval $[1-\varepsilon, 1+\varepsilon]$ where ε is small (almost zero).

If, moreover, the $n \times n$ linear system $A\mathbf{x} = \mathbf{b}$ can be seen as one of a sequence of increasing order linear systems, with the property that $\forall\, \varepsilon > 0\ \exists\, k_\varepsilon$, n_ε such that for all $n > n_\varepsilon$ outside $[1-\varepsilon, 1+\varepsilon]$ fall no more than n_ε eigenvalues of A, then (bound) allows to prove the superlinear convergence of CG.

(Note that in general CG has a linear rate of convergence, as a consequence of (wkbound)).

Let us prove these assertions, by evaluating $\max_{\lambda \in S} |p_q(\lambda)|$.

$$\begin{aligned} \max_{\lambda \in S} |p_q(\lambda)| &= \max_{\lambda \in [1-\varepsilon, 1+\varepsilon]} |p_q(\lambda)| \\ &\leq (\max_{\ldots} \prod_{\lambda_j^\varepsilon} \left|1 - \frac{\lambda}{\lambda_j^\varepsilon}\right|)(\max_{\ldots} \left|\frac{T_{q-r_\varepsilon}((1-\lambda)/\varepsilon)}{T_{q-r_\varepsilon}(1/\varepsilon)}\right|) \\ &= (\max_{\ldots} \prod_{\lambda_j^\varepsilon} \left|1 - \frac{\lambda}{\lambda_j^\varepsilon}\right|)\frac{1}{T_{q-r_\varepsilon}(1/\varepsilon)}. \end{aligned}$$

Now first notice that

$$T_{q-r_\varepsilon}\left(\frac{1}{\varepsilon}\right) = T_{q-r_\varepsilon}\left(\frac{\frac{1+\varepsilon}{1-\varepsilon} + 1}{\frac{1+\varepsilon}{1-\varepsilon} - 1}\right) > \frac{1}{2}\left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} + 1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} - 1}\right)^{q-r_\varepsilon}.$$

Then denote by $\hat{\lambda}_j^\varepsilon$ those eigenvalues λ_j^ε satisfying the inequalities

$$\lambda_j^\varepsilon < 1 - \varepsilon, \quad \lambda_j^\varepsilon < \frac{1}{2}(1+\varepsilon)$$

and observe that

$$\begin{aligned} \max_{\lambda \in [1-\varepsilon, 1+\varepsilon]} \prod_{\lambda_j^\varepsilon} \left|1 - \frac{\lambda}{\lambda_j^\varepsilon}\right| &\leq \max_{\ldots} \prod_{\hat{\lambda}_j^\varepsilon} \left|1 - \frac{\lambda}{\lambda_j^\varepsilon}\right| \\ &= \prod_{\hat{\lambda}_j^\varepsilon} \left(\frac{1+\varepsilon}{\lambda_j^\varepsilon} - 1\right). \end{aligned}$$

So, we have

$$\begin{aligned} \max_{\lambda \in S} |p_q(\lambda)| &\leq \prod_{\hat{\lambda}_j^\varepsilon} \left(\frac{1+\varepsilon}{\lambda_j^\varepsilon} - 1\right) 2\left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} - 1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} + 1}\right)^{q-r_\varepsilon} \\ &\leq 2\left(\frac{1+\varepsilon}{\min \lambda(A)} - 1\right)^{\#\hat{\lambda}_j^\varepsilon} \left(\frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} - 1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} + 1}\right)^{q-r_\varepsilon} \\ &\approx \left(\frac{1+\varepsilon}{\min \lambda(A)} - 1\right)^{\#\hat{\lambda}_j^\varepsilon} \frac{\varepsilon^q}{\varepsilon^{r_\varepsilon} 2^{q-r_\varepsilon - 1}}, \end{aligned}$$

where in the latter approximation we have used the following Taylor expansion

$$f(\varepsilon) = \frac{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} - 1}{\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} + 1} = \frac{\varepsilon}{2} + \frac{\varepsilon^2}{2} f''(0) + \dots .$$