$$M = \begin{bmatrix} 1 & 1/2 & 1/3 & . & . & \frac{1}{n} \\ 1/2 & 1/3 & . & . & \frac{1}{n} & \frac{1}{n+1} \\ 1/3 & . & . & . & . & . \\ . & . & . & . & . & . \\ . & \frac{1}{n} & . & . & . & . \\ \frac{1}{n} & \frac{1}{n+1} & . & . & . & \frac{1}{2n-1} \end{bmatrix}$$ Matrice di Hilbert

$$M \in \mathbb{C}^{r\times s},\ N \in \mathbb{C}^{s\times t} \quad \Rightarrow \quad (MN)^H = N^H M^H,\ (MN)^T = N^T M^T$$

$$M, N \in \mathbb{C}^{n\times n} \text{ non singular} \quad \Rightarrow \quad (MN)^{-1} = N^{-1} M^{-1}$$

$$M \in \mathbb{C}^{r\times s},\ N \in \mathbb{C}^{s\times r},\ MN \text{ non singular} \quad \Rightarrow \quad (MN)^{-1} = ?$$

$$A\mathbf{x} = \lambda\mathbf{x},\ \mathbf{x} \neq \mathbf{0} \quad \Rightarrow \quad \lambda = \frac{\mathbf{x}^H A\mathbf{x}}{\mathbf{x}^H\mathbf{x}},\ \overline{\lambda} = \frac{\mathbf{x}^H A^H\mathbf{x}}{\mathbf{x}^H\mathbf{x}}$$

$$A^H = A,\ -A,\ A^{-1} \quad \Rightarrow \quad \overline{\lambda} = \lambda,\ -\lambda,\ \lambda^{-1}$$

Let x_k, $k = 0, \dots, N-1$, be N distinct real numbers, and $f(x_k)$ the values of a function f (that might be unknown) in such x_k. Assume that ϕ_i, $i = 0, \dots, n-1$, are n simple linearly independent functions defined on a subset of $\mathbb{R}$ including the x_k. The aim is to investigate when the following (error) function

$$E(\mathbf{a}) = \sum_{k=0}^{N-1} \Big(\sum_{i=0}^{n-1} a_i \phi_i(x_k) - f(x_k) \Big)^2 = c - 2\mathbf{b}^T\mathbf{a} + \mathbf{a}^T M \mathbf{a}, \quad \mathbf{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_{n-1} \end{bmatrix} \in \mathbb{R}^n,$$

$$c = \sum_{k=0}^{N-1} f(x_k)^2,\ b_i = \sum_{k=0}^{N-1} \phi_i(x_k) f(x_k),\ M_{ij} = \sum_{k=0}^{N-1} \phi_i(x_k)\phi_j(x_k),\ i,j = 0, \dots, n-1,\ \mathbf{a}^T M \mathbf{a} = \sum_{k=0}^{N-1} (\sum_{i=0}^{n-1} a_i \phi_i(x_k))^2$$

has a unique minimum $\hat{\mathbf{a}}$, and therefore it is well defined in the set $\{\sum_{i=0}^{n-1} a_i \phi_i(x) : a_i \in \mathbb{R}\}$ the best approximation $\sum_{i=0}^{n-1} \hat{a}_i \phi_i(x)$ of the table $(x_k, f(x_k))$, $k = 0, \dots, N-1$, nel senso dei minimi quadrati.

The matrix M is $n \times n$ real symmetric positive semi-definite, $M = Y^T Y$, $Y_{kj} = \phi_j(x_k)$, Y $N \times n$. M is non singular (i.e. the system $\frac{1}{2}\nabla f(\mathbf{a}) = M\mathbf{a} - \mathbf{b} = Y^T(Y\mathbf{a} - f(\mathbf{x})) = \mathbf{0}$ has a unique solution) if and only if it is positive definite [this is true for any positive semi-definite matrix]. $\mathbf{a}^T M \mathbf{a} = 0$ if and only if $Y\mathbf{a} = \mathbf{0}$, $[Y\mathbf{a}]_k = \sum_{i=0}^{n-1} a_i \phi_i(x_k)$, $k = 0, \dots, N-1$.

Now, if $N < n$, then certainly the system $Y\mathbf{a} = \mathbf{0}$ has non null vector solutions $\mathbf{a}$, i.e. $\exists \mathbf{a} \neq \mathbf{0}$ such that $\mathbf{a}^T M \mathbf{a} = 0$ and M is singular. So, assume from now on that $N \geq n$. Then $Y\mathbf{a} = \mathbf{0}$ implies $\mathbf{a} = \mathbf{0}$ if and only if the columns of the high rectangular matrix

$$Y = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & . & \phi_{n-1}(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & . & \phi_{n-1}(x_1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(x_{N-1}) & \phi_1(x_{N-1}) & . & \phi_{n-1}(x_{N-1}) \end{bmatrix}$$

are linearly independent.

Note that the linearly independence of the ϕ_i is a necessary condition for the columns of Y to be linearly independent. However, it can happen that, due of an unlucky choice of the x_k, the columns of Y are linearly dependent even if the ϕ_i are independent.

Thus *if the ϕ_i, $i = 0, \dots, n-1$, are linearly independent, and, together with $N \geq n$ distinct real numbers x_k, are such that the columns of the $N \times n$ matrix Y, $Y_{kj} = \phi_j(x_k)$, are linearly independent, then $E(\mathbf{a})$ is minimum for $\mathbf{a} = \hat{\mathbf{a}} = M^{-1}\mathbf{b} = (Y^TY)^{-1}Y^T f(\mathbf{x})$ and $E(\hat{\mathbf{a}}) = c - 2\mathbf{b}^T\hat{\mathbf{a}} + \hat{\mathbf{a}}^T M \hat{\mathbf{a}} = c - \mathbf{b}^T M^{-1}\mathbf{b} = \|f(\mathbf{x})\|^2 - f(\mathbf{x})^T Y (Y^TY)^{-1} Y^T f(\mathbf{x})$. If moreover $N = n$, then $\hat{\mathbf{a}} = M^{-1}\mathbf{b} = Y^{-1} f(\mathbf{x})$ and $E(\hat{\mathbf{a}}) = c - 2\mathbf{b}^T\hat{\mathbf{a}} + \hat{\mathbf{a}}^T M \hat{\mathbf{a}} = \|f(\mathbf{x})\|^2 - f(\mathbf{x})^T f(\mathbf{x}) = 0$.*

If the ϕ_i are degree-i poynomials then the columns of Y are linearly independent, and $\sum_{i=0}^{n-1} \hat{a}_i \phi_i(x)$ is the best degree$\leq n-1$ polynomial approximating $(x_k, f(x_k))$, $k = 0, \dots, N-1$. The ϕ_i can be chosen such that $Y^TY = I$, so that $\hat{\mathbf{a}} = Y^T f(\mathbf{x})$ (i.e. no system needs to be solved to compute $\hat{\mathbf{a}}$).

Theorem: Let M be a $n \times n$ matrix with complex entries, i.e. $M \in \mathbb{C}^{n\times n}$. Then the homogeneous linear system $M\mathbf{y} = \mathbf{0}$ has solutions $\mathbf{y} \neq \mathbf{0}$ if and only if $\det(M) = 0$.

Proof. Assume that $M\mathbf{y} = \mathbf{0}$ has solutions $\mathbf{y} \neq \mathbf{0}$. We want to prove that $\det(M) = 0$. Suppose that it is not true, i.e. that $\det(M) \neq 0$; in this case M is invertible and the vector equation $M\mathbf{y} = \mathbf{0}$ implies $M^{-1}M\mathbf{y} = M^{-1}\mathbf{0} = \mathbf{0} \Rightarrow \mathbf{y} = \mathbf{0}$, i.e. the only solution of $M\mathbf{y} = \mathbf{0}$ is the null vector, which is a contradiction with the hypothesis.

Now assume that $\det(M) = 0$. We want to prove that $M\mathbf{y} = \mathbf{0}$ has solutions $\mathbf{y} \neq \mathbf{0}$. Since $\det(M) = 0$, we can say that the columns of M, $M\mathbf{e}_1$, $M\mathbf{e}_2$, ..., $M\mathbf{e}_n$, are linearly dependent, thus there exist complex numbers α_1, α_2, ..., α_n not all zero such that $\alpha_1 M\mathbf{e}_1 + \alpha_2 M\mathbf{e}_2 + \ldots + \alpha_n M\mathbf{e}_n = \mathbf{0}$, but this is like to say that

$$M \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ \alpha_n \end{bmatrix} = [\, M\mathbf{e}_1 \; M\mathbf{e}_2 \; \ldots \; M\mathbf{e}_n \,] \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ \alpha_n \end{bmatrix} = \mathbf{0}, \quad \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ \alpha_n \end{bmatrix} \neq \mathbf{0}$$

that is, the thesis.

$\lambda \in \mathbb{C}$ is an *eigenvalue* of $A \in \mathbb{C}^{n\times n}$ if and only if $\exists\, \mathbf{x} \in \mathbb{C}^n$ $\mathbf{x} \neq \mathbf{0}$ such that $A\mathbf{x} = \lambda\mathbf{x}$ if and only if $\exists\, \mathbf{x} \neq \mathbf{0}$ such that $(\lambda I - A)\mathbf{x} = \mathbf{0}$ if and only if (by the above Theorem) $\det(\lambda I - A) = 0$.

If $\lambda \in \mathbb{C}$ is an eigenvalue of A then the set of all $\mathbf{x} \in \mathbb{C}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$ (*eigenvectors* of A associated with λ) is called *eigenspace* of A associated with λ.

So the eigenvalues of $A \in \mathbb{C}^{n\times n}$ are the roots of the *characteristic equation* $\det(\lambda I - A) = 0$. One can observe that $\det(\lambda I - A)$ is a degree n monic polynomial in λ whose coefficients depend on the entries of A; in particular, it can be proved that the coefficients of λ^{n-1} and of λ^0 are $-\sum_{i=1}^n a_{ii}$ and $(-1)^n \det(A)$, respectively. So the *characteristic polynomial* has the following form:

$$\det(\lambda I - A) = \lambda^n - (\sum_{i=1}^n a_{ii})\lambda^{n-1} + \ldots + (-1)^{n-1}(...)\lambda + (-1)^n(\det(A)).$$

Since, by the fundamental theorem of algebra, any degree n polynomial with coefficients in $\mathbb{C}$ has exactly n roots in $\mathbb{C}$, it follows that the eigenvalues of A are n, and we can call them $\lambda_1 = \lambda_1(A)$, $\lambda_2 = \lambda_2(A)$, ..., $\lambda_n = \lambda_n(A)$.

Important remark: by the definition of the $\lambda_i(A)$, the equality

$$\det(\lambda I - A) = \lambda^n - (\sum_{i=1}^n a_{ii})\lambda^{n-1} + \ldots + (-1)^{n-1}(..)\lambda + (-1)^n(\det(A)) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_{n-1})(\lambda - \lambda_n)$$

must hold. Such equality implies the following two important relations between the entries and the eigenvalues of A:

$$\sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i, \qquad \det(A) = \prod_{i=1}^n \lambda_i.$$

For example, if $A \in \mathbb{C}^{2\times 2}$: $\det(\lambda I - A) = \lambda^2 - (a_{11} + a_{22})\lambda + (\det(A)) = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2$ implies $a_{11} + a_{22} = \lambda_1 + \lambda_2$, $\det(A) = \lambda_1\lambda_2$.

If $A \in \mathbb{C}^{3\times 3}$: $\det(\lambda I - A) = \lambda^3 - (a_{11} + a_{22} + a_{33})\lambda^2 + (...)\lambda - (\det(A)) = (\lambda - \lambda_1)(\lambda - \lambda_2)(\lambda - \lambda_3) = \lambda^3 - (\lambda_1 + \lambda_2 + \lambda_3)\lambda^2 + (...)\lambda - \lambda_1\lambda_2\lambda_3$ implies $a_{11} + a_{22} + a_{33} = \lambda_1 + \lambda_2 + \lambda_3$, $\det(A) = \lambda_1\lambda_2\lambda_3$.

$[\mathbf{r}_1 \ \dots \ \mathbf{r}_n] = R = VT$ with $V = [\mathbf{v}_1 \ \dots \ \mathbf{v}_n]$ unitary and T upper triangular invertible $\Rightarrow$

$$V = RT^{-1} \ \Rightarrow \ V^H BV = T^{-H} R^H BRT^{-1} \ \Rightarrow$$

$$\mathbf{v}_k^H B\mathbf{v}_k = [V^H BV]_{kk} = [T^{-H} R^H BRT^{-1}]_{kk}$$

$$= \sum_{r,s} [T^{-H}]_{kr} [R^H BR]_{rs} [T^{-1}]_{sk} = \sum_{r=1}^{k} [T^{-H}]_{kr} [R^H BR]_{rr} [T^{-1}]_{rk}$$

(because the $\mathbf{r}_i$ are B-conjugate and T^{-1} is upper triangular)

$$= \sum_{r=1}^{k} \overline{[T^{-1}]_{rk}} [R^H BR]_{rr} [T^{-1}]_{rk} = \sum_{r=1}^{k} |[T^{-1}]_{rk}|^2 [R^H BR]_{rr}$$

$$= |[T^{-1}]_{kk}|^2 [R^H BR]_{kk} + \sum_{r=1}^{k-1} |[T^{-1}]_{rk}|^2 [R^H BR]_{rr}$$

$$= [R^H BR]_{kk} + (|[T^{-1}]_{kk}|^2 - 1)[R^H BR]_{kk} + \sum_{r=1}^{k-1} |[T^{-1}]_{rk}|^2 [R^H BR]_{rr}$$

$$= [R^H BR]_{kk} + \frac{1 - |T_{kk}|^2}{|T_{kk}|^2} [R^H BR]_{kk} + \sum_{r=1}^{k-1} |[T^{-1}]_{rk}|^2 [R^H BR]_{rr}$$

But $T_{kk} = [V^H R]_{kk} = \mathbf{v}_k^H \mathbf{r}_k$, thus $|T_{kk}| \leq \|\mathbf{v}_k\| \|\mathbf{r}_k\| = \|\mathbf{r}_k\| \leq 1$ (here the hypothesis $\|\mathbf{r}_k\| \leq 1$ is used) $\Rightarrow \ 1 - |T_{kk}|^2 \geq 0 \Rightarrow \mathbf{v}_k^H B\mathbf{v}_k \geq \mathbf{r}_k^H B\mathbf{r}_k$.

$\mathbf{r}_1 = T_{11}\mathbf{v}_1$, $\|\mathbf{r}_1\| \leq 1 \Rightarrow 1 \geq \|\mathbf{r}_1\| = \|T_{11}\mathbf{v}_1\| = |T_{11}| \|\mathbf{v}_1\| = |T_{11}| \Rightarrow \mathbf{v}_1^H B\mathbf{v}_1 = \frac{1}{|T_{11}|^2} \mathbf{r}_1^H B\mathbf{r}_1 \geq \mathbf{r}_1^H B\mathbf{r}_1$

Let $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$ be vectors of $\mathbb{C}^n$ such that $\mathbf{r}_i^H B \mathbf{r}_j = 0$, $i \neq j$, where B is a Hermitian positive definite $n \times n$ matrix. Assume, moreover, that $\mathbf{r}_i^H \mathbf{r}_i = 1 \; \forall \, i$. Apply Gram-Schmidt to $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$:

$$\mathbf{v}_1 = \frac{\mathbf{r}_1}{\|\mathbf{r}_1\|}, \; \mathbf{v}_2 = \frac{\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1}{\|\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1\|}, \; \ldots, \; \mathbf{v}_k = \frac{\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j}{\|\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j\|}, \ldots$$

Note that

$$\mathbf{v}_1^H B \mathbf{v}_1 = \frac{\mathbf{r}_1^H B \mathbf{r}_1}{\mathbf{r}_1^H \mathbf{r}_1} = \mathbf{r}_1^H B \mathbf{r}_1,$$

where the last equality follows from the hypothesis $\mathbf{r}_1^H \mathbf{r}_1 = 1$. Thus $\underline{\mathbf{r}_1^H B \mathbf{r}_1 = \mathbf{v}_1^H B \mathbf{v}_1}$.
Note also that, by the hypothesis $\mathbf{r}_2^H B \mathbf{r}_1 = 0$, we have

$$\mathbf{v}_2^H B \mathbf{v}_2 = \frac{(\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1)^H B (\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1)}{(\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1)^H (\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1)} = \frac{(\mathbf{r}_2^H - \overline{(\mathbf{v}_1^H \mathbf{r}_2)}\mathbf{v}_1^H) B (\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1)}{(\mathbf{r}_2^H - \overline{(\mathbf{v}_1^H \mathbf{r}_2)}\mathbf{v}_1^H)(\mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{v}_1)}$$

$$= \frac{\mathbf{r}_2^H B \mathbf{r}_2 - \mathbf{v}_1^H \mathbf{r}_2 \mathbf{r}_2^H B \mathbf{v}_1 - \overline{(\mathbf{v}_1^H \mathbf{r}_2)}\mathbf{v}_1^H B \mathbf{r}_2 + \overline{(\mathbf{v}_1^H \mathbf{r}_2)}\mathbf{v}_1^H \mathbf{r}_2 \mathbf{v}_1^H B \mathbf{v}_1}{\mathbf{r}_2^H \mathbf{r}_2 - (\mathbf{v}_1^H \mathbf{r}_2)\mathbf{r}_2^H \mathbf{v}_1 - \overline{(\mathbf{v}_1^H \mathbf{r}_2)}\mathbf{v}_1^H \mathbf{r}_2 + (\mathbf{v}_1^H \mathbf{r}_2)\overline{(\mathbf{v}_1^H \mathbf{r}_2)}\mathbf{v}_1^H \mathbf{v}_1} = \frac{\mathbf{r}_2^H B \mathbf{r}_2 + |\mathbf{v}_1^H \mathbf{r}_2|^2 \mathbf{v}_1^H B \mathbf{v}_1}{\mathbf{r}_2^H \mathbf{r}_2 - |\mathbf{v}_1^H \mathbf{r}_2|^2}$$

$$= \frac{\mathbf{r}_2^H B \mathbf{r}_2 (1 - |\mathbf{v}_1^H \mathbf{r}_2|^2) + \mathbf{r}_2^H B \mathbf{r}_2 |\mathbf{v}_1^H \mathbf{r}_2|^2 + |\mathbf{v}_1^H \mathbf{r}_2|^2 \mathbf{v}_1^H B \mathbf{v}_1}{\mathbf{r}_2^H \mathbf{r}_2 - |\mathbf{v}_1^H \mathbf{r}_2|^2} = \mathbf{r}_2^H B \mathbf{r}_2 + \frac{\mathbf{r}_2^H B \mathbf{r}_2 |\mathbf{v}_1^H \mathbf{r}_2|^2 + |\mathbf{v}_1^H \mathbf{r}_2|^2 \mathbf{v}_1^H B \mathbf{v}_1}{1 - |\mathbf{v}_1^H \mathbf{r}_2|^2}$$

where the last equality follows from the hypothesis $\mathbf{r}_2^H \mathbf{r}_2 = 1$. Thus $\underline{\mathbf{r}_2^H B \mathbf{r}_2 \leq \mathbf{v}_2^H B \mathbf{v}_2}$.
At the generic step, by using the hypothesis $\mathbf{r}_k^H B \mathbf{r}_j = 0$, $j = 1, \ldots, k-1$, note that

$$\mathbf{v}_k^H B \mathbf{v}_k = \frac{(\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)^H B (\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)}{(\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)^H (\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)} = \frac{(\mathbf{r}_k^H - \sum_{j=1}^{k-1} \overline{(\mathbf{v}_j^H \mathbf{r}_k)}\mathbf{v}_j^H) B (\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)}{(\mathbf{r}_k^H - \sum_{j=1}^{k-1} \overline{(\mathbf{v}_j^H \mathbf{r}_k)}\mathbf{v}_j^H)(\mathbf{r}_k - \sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)}$$

$$= \frac{\mathbf{r}_k^H B \mathbf{r}_k + (\sum_{j=1}^{k-1} \overline{(\mathbf{v}_j^H \mathbf{r}_k)}\mathbf{v}_j^H) B (\sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)}{\mathbf{r}_k^H \mathbf{r}_k - \sum_{j=1}^{k-1} |\mathbf{v}_j^H \mathbf{r}_k|^2}$$

$$= \frac{\mathbf{r}_k^H B \mathbf{r}_k (1 - \sum_{j=1}^{k-1} |\mathbf{v}_j^H \mathbf{r}_k|^2) + \mathbf{r}_k^H B \mathbf{r}_k \sum_{j=1}^{k-1} |\mathbf{v}_j^H \mathbf{r}_k|^2 + (\sum_{j=1}^{k-1} \overline{(\mathbf{v}_j^H \mathbf{r}_k)}\mathbf{v}_j^H) B (\sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)}{\mathbf{r}_k^H \mathbf{r}_k - \sum_{j=1}^{k-1} |\mathbf{v}_j^H \mathbf{r}_k|^2}$$

$$= \mathbf{r}_k^H B \mathbf{r}_k + \frac{\mathbf{r}_k^H B \mathbf{r}_k \sum_{j=1}^{k-1} |\mathbf{v}_j^H \mathbf{r}_k|^2 + (\sum_{j=1}^{k-1} \overline{(\mathbf{v}_j^H \mathbf{r}_k)}\mathbf{v}_j^H) B (\sum_{j=1}^{k-1}(\mathbf{v}_j^H \mathbf{r}_k)\mathbf{v}_j)}{1 - \sum_{j=1}^{k-1} |\mathbf{v}_j^H \mathbf{r}_k|^2}$$

where the last equality follows from the hypothesis $\mathbf{r}_k^H \mathbf{r}_k = 1$. Thus $\underline{\mathbf{r}_k^H B \mathbf{r}_k \leq \mathbf{v}_k^H B \mathbf{v}_k}$.

So we obtain the matrix identity $[\mathbf{r}_1 \cdots \mathbf{r}_n] = [\mathbf{v}_1 \cdots \mathbf{v}_n] T$, with T upper triangular, $T_{ij} = \mathbf{v}_i^H \mathbf{r}_j$, $i < j$, $T_{ii} = \mathbf{v}_i^H \mathbf{r}_i = \|\mathbf{r}_i - \sum_{j=1}^{i-1}(\mathbf{v}_j^H \mathbf{r}_i)\mathbf{v}_j\| = \sqrt{1 - \sum_{j=1}^{i-1} |\mathbf{v}_j^H \mathbf{r}_i|^2}$, $T_{ij} = 0$ $i > j$.

Now apply any other QR orthonormalization algorithm to $R = [\mathbf{r}_1 \cdots \mathbf{r}_n]$, and call $\hat{\mathbf{v}}_i$ the orthonormal vectors so generated, thus $R = [\hat{\mathbf{v}}_1 \cdots \hat{\mathbf{v}}_n]\hat{T}$ for some upper triangular $\hat{T}$. But then it must exist D diagonal unitary such that $[\hat{\mathbf{v}}_1 \cdots \hat{\mathbf{v}}_n] = [\mathbf{v}_1 \cdots \mathbf{v}_n] D$, i.e. θ_k for which $\hat{\mathbf{v}}_k = e^{\mathbf{i}\theta_k} \mathbf{v}_k$; this implies that $\hat{\mathbf{v}}_k^H B \hat{\mathbf{v}}_k = \mathbf{v}_k^H B \mathbf{v}_k \geq \mathbf{r}_k^H B \mathbf{r}_k$.

Definitions:

A matrix $A \in \mathbb{C}^{n\times n}$ is a square array with n^2 entries a_{ij}, $i,j = 1,\ldots,n$, that are real or complex numbers. Associated with a matrix A, there are its *eigenvalues*, which are complex numbers

$$\lambda \in \mathbb{C} \text{ such that } A\mathbf{x} = \lambda\mathbf{x}, \text{ for some non null vector } \mathbf{x} \in \mathbb{C}^n.$$

(Note that if λ is not real and A is real then $\mathbf{x}$ must be not real.) The set of all *eigenvectors* $\mathbf{x}$ of λ, satisfying the above equality, forms a vector space (subspace of $\mathbb{C}^n$), known as the *eigenspace* of the eigenvalue λ of A. Note that such space is invariant under the action of A. *Given two distinct eigenvalues of A, λ_1 and λ_2, any eigenvector of λ_1 is linearly independent with any eigenvector of* λ_2, in fact, if $A\mathbf{x}_1 = \lambda_1\mathbf{x}_1$, $\mathbf{x}_1 \neq \mathbf{0}$, $A\mathbf{x}_2 = \lambda_2\mathbf{x}_2$, $\mathbf{x}_2 \neq \mathbf{0}$, and $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 = \mathbf{0}$, then $\mathbf{0} = p(A)(\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2) = \alpha_1 p(\lambda_1)\mathbf{x}_1 + \alpha_2 p(\lambda_2)\mathbf{x}_2$, for all polynomials p [1]; in particular, for $p(x) = (x-\lambda_2)/(\lambda_1-\lambda_2)$ and $p(x) = (x-\lambda_1)/(\lambda_2-\lambda_1)$, choices allowed if $\lambda_1 \neq \lambda_2$, we obtain that $\alpha_1\mathbf{x}_1 = \mathbf{0}$ and $\alpha_2\mathbf{x}_2 = \mathbf{0}$, which imply $\alpha_1 = 0$ and $\alpha_2 = 0$, respectively.

In the eigenspace of λ one can choose a set of linearly independent vectors spanning the eigenspace. This can be repeated for each distinct eigenvalue of A. By collecting all such sets, one can form a rectangular $n \times m$ matrix R, $n \geq m$, whose columns are linearly independent, such that

$$AR = RD, \quad D \text{ diagonal } m \times m, \quad D_{ii} \in \{\text{the distinct eigenvalues of } A\}.$$

If m turns out to be equal to n, then R is square $n \times n$ and invertible (so its columns form a basis for $\mathbb{C}^n$), D is $n \times n$ with eigenvalues of A as diagonal entries (they are all the eigenvalues of A, why?), and the identity $AR = RD$ can be rewritten as $R^{-1}AR = D$, in other words, A turns out to be *diagonalizable* by a similarity transform.

If $m < n$, then it is not possible to diagonalize A by a similarity transform (why?); anyway R can be completed, involving suitable vectors as new columns, so to become an invertible square $n \times n$ matrix χ such that $\chi^{-1}A\chi = J$, with J block diagonal with diagonal blocks of type

$$\mu_k I_{s_k} + Z_{s_k}^T = \left.\begin{bmatrix} \mu_k & 1 & 0 & . & 0 \\ 0 & \mu_k & 1 & . & . \\ . & . & . & . & . \\ 0 & . & . & . & 1 \\ 0 & . & . & 0 & \mu_k \end{bmatrix}\right\} s_k \ ,$$

$\mu_k \in \{\text{eigenvalues of } A\}$, where at least one of them is of order at least 2.

Since the equation $A\mathbf{x} = \lambda\mathbf{x}$ is equivalent to the homogeneous linear system $(\lambda I - A)\mathbf{x} = \mathbf{0}$, the eigenvalues of A can be obtained by solving the *characteristic equation* $\det(\lambda I - A) = 0$, since the latter is a NASC (Necessary and Sufficient Condition) on λ for the existence of non null solutions $\mathbf{x}$ of $(\lambda I - A)\mathbf{x} = \mathbf{0}$. If one writes the matrix $\lambda I - A$, then it is clear that $\det(\lambda I - A)$ is a monic polynomial in λ of degree n whose coefficients are functions of the entries a_{ij} of A, and are real whenever the a_{ij} are real. Thus the set $\sigma(A)$ of the eigenvalues λ of A coincide with the set of the n roots $\lambda_1, \lambda_2, \ldots, \lambda_n$ of the following algebraic equation:

$$\det(\lambda I - A) = \lambda^n - (\sum_i a_{ii})\lambda^{n-1} + \ldots + (-1)^n \det(A) = 0$$

[1] $A\mathbf{x} = \lambda\mathbf{x} \Rightarrow A^k\mathbf{x} = \lambda^k\mathbf{x} \Rightarrow \sum_k \alpha_k A^k\mathbf{x} = \sum_k \alpha_k\lambda^k\mathbf{x} \Rightarrow p(A)\mathbf{x} = p(\lambda)\mathbf{x}$ with $p(t) = \sum_k \alpha_k t^k$

(the proof of the expressions of the coefficients of λ^0 and of λ^{n-1} is omitted). As a consequence, if the a_{ij} are real, then $\lambda \in \sigma(A) \Rightarrow \overline{\lambda} \in \sigma(A)$ (why?), i.e. the set $\sigma(A)$ is closed under conjugation. Note that the representation of the *characteristic polynomial* $p_A(\lambda) = \det(\lambda I - A)$ in terms of its zeros, $\det(\lambda I - A) = \Pi_i(\lambda - \lambda_i)$, implies the following two important identities, $\sum_i a_{ii} = \sum_i \lambda_i$ and $\det(A) = \Pi_i \lambda_i$, relating the eigenvalues of A with its entries. For example, the latter implies that a matrix A is singular (has zero determinant) if and only if at least one of its eigenvalues is zero.

We have seen that associated with any matrix A there is a monic degree n polynomial, i.e. the characteristic polynomial $p_A(\lambda)$. Viceversa, consider any monic degree n polynomial $x^n + a_{n-1}x^{n-1} + \ldots + a_2x^2 + a_1x + a_0$. Is it possible to write a $n \times n$ matrix A with such polynomial as characteristic polynomial? The answer is yes, just take the so called Frobenius (companion) matrix associated with the polynomial

$$A = \begin{bmatrix} 0 & 1 & 0 & . & 0 \\ . & 0 & 1 & . & . \\ . & . & . & . & 0 \\ 0 & 0 & . & 0 & 1 \\ -a_0 & -a_1 & -a_2 & . & -a_{s-1} \end{bmatrix}, \quad \text{write} \quad \lambda I - A = \begin{bmatrix} \lambda & -1 & 0 & . & 0 \\ 0 & \lambda & -1 & . & . \\ . & . & . & . & 0 \\ 0 & 0 & . & \lambda & -1 \\ a_0 & a_1 & a_2 & . & \lambda + a_{s-1} \end{bmatrix},$$

and calculate $\det(\lambda I - A)$ by applying Laplace rule to the first column of $\lambda I - A$. For example

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 \end{bmatrix}, \quad \lambda I - A = \begin{bmatrix} \lambda & -1 & 0 \\ 0 & \lambda & -1 \\ a_0 & a_1 & \lambda + a_2 \end{bmatrix}, \quad \Rightarrow$$

$$\det(\lambda I - A) = \lambda(\lambda(\lambda + a_2) + a_1) + a_0(-1)(-1) = \lambda^3 + \lambda^2 a_2 + \lambda a_1 + a_0.$$

Thus the computation of the roots of an algebraic equation is equivalent to the computation of the eigenvalues of a matrix.

Instead of computing the eigenvalues of a $n \times n$ matrix A, it is often sufficient to localize them, i.e. find some region of the complex field including all or some of them. In particular, Gershgorin theorem allows one to localize all of them in the union of n circles easily defined from the entries of A, in fact it states that

$$\lambda \in \mathbb{C} \ : A\mathbf{x} = \lambda\mathbf{x}, \, \mathbf{x} \neq \mathbf{0}, \quad \Rightarrow \quad \lambda \in \cup_{i=1}^n K_i, \quad K_i = \{z \in \mathbb{C} : \, |z - a_{ii}| \leq \sum_{j=1, \, j \neq i}^{n} |a_{ij}|\}.$$

When A is irreducible, a more precise assertion holds: $\lambda \in \sigma(A) \Rightarrow$ either λ is in the inner part of at least one of the Gershgorin circles, or it is on the border of each Gershgorin circle. Recall that a matrix is reducible if there exists $\mathcal{I} \subset \{1, 2, \ldots, n\}$, $\mathcal{I} \neq \{1, 2, \ldots, n\}$, $\mathcal{I} \neq \emptyset$, such that $a_{ij} = 0$ $\forall \, i \in \mathcal{I}$, $j \in \{1, 2, \ldots, n\} \backslash \mathcal{I}$, or, equivalently, if there is a permutation matrix P for which $P^T A P$ is an upper (or lower) 2×2 block triangular matrix with square diagonal blocks.

Triangular matrices have several interesting properties. Let us see some of them. Assume that $A \in \mathbb{C}^{n \times n}$ is such that $a_{ij} = 0$ for all $i > j$ / $i < j$, i.e. that A is upper/lower triangular. Then (i) the eigenvalues of A are its diagonal entries; (ii) A is invertible if and only if there is no zero on its diagonal; (iii) the matrix A^{-1} (when defined) is upper/lower triangular like A and $[A^{-1}]_{ii} = \frac{1}{a_{ii}}$;

(iv) if B is any other upper/lower triangular matrix, then AB and BA are upper/lower triangular matrices; (v) the linear system $A\mathbf{x} = \mathbf{b}$ can be easily solved by the backward/forward substitution algorithm with $\frac{n(n+1)}{2}$ multiplicative operations.

A $n \times n$ matrix T is said to be *Toeplitz* if the entries on each diagonal of T are all equal. A *Hankel* matrix H is obtained by reversing the columns (or the rows) of a Toeplitz matrix. Such operation is equivalent to a multiplication on the right (or on the left) by the reversion matrix J. For example

$$T = \begin{bmatrix} c & d & e \\ b & c & d \\ a & b & c \end{bmatrix}, \quad H = TJ = \begin{bmatrix} e & d & c \\ d & c & b \\ c & b & a \end{bmatrix}, \quad H = JT = \begin{bmatrix} a & b & c \\ b & c & d \\ c & d & e \end{bmatrix}, \quad J = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Multiply a matrix A by a vector, compute the eigenvalues of A, solve a linear system $A\mathbf{x} = \mathbf{b}$, are all operations that become simpler when A has *Toeplitz or Hankel structure* (...). A Hankel matrix

$$H = \begin{bmatrix} a & b & c \\ b & c & d \\ c & d & e \end{bmatrix},$$

is Hermitian if and only if it is real.

Triangular Toeplitz matrices have even more interesting properties than triangular matrices. Let $T \in \mathbb{C}^{n \times n}$ be a lower/upper triangular Toeplitz matrix with the parameters $a_0, a_1, \ldots, a_{n-1}$ on its first column/row. (i) T is a polynomial in the matrix Z / Z^T, in fact $T = p(Z)$ / $p(Z^T)$ for $p(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1}$. (ii) if T' is any other lower/upper triangular Toeplitz matrix, then TT' is equal to $T'T$ and is yet a lower/upper triangular Toeplitz matrix; (iii) the matrix T^{-1} (when defined) is lower/upper triangular Toeplitz like T; (iv) the product $T * \mathbf{v}$, $\mathbf{v} \in \mathbb{C}^n$, can be computed with $O(n \log_2 n)$ arithmetic operations; (v) the linear system $T\mathbf{x} = \mathbf{b}$ can be solved with $O(n \log_2 n)$ arithmetic operations, computing the first column/row of T^{-1}, and then using (iii) and (iv).

I numeri di Fibonacci sono talmente tanti, che non finiscono mai

I primi due numeri di Fibonacci sono 0 e 1. Il terzo si ottiene sommandoli:

$0 + 1 = 1$, $\quad 1 =$ terzo numero di Fibonacci.

Il quarto si ottiene sommando al secondo addendo nella somma qui sopra il risultato della somma. Quindi:

$1 + 1 = 2$, $\quad 2 =$ quarto numero di Fibonacci.

Il quinto si ottiene sommando al secondo addendo nella somma qui sopra il risultato della somma. Quindi:

$1 + 2 = 3$, $\quad 3 =$ quinto numero di Fibonacci.

I successivi numeri (il sesto, il settimo, ...) si ottengono con la stessa regola:

$2 + 3 = 5$, $\quad 5 =$ sesto numero di Fibonacci,

$3 + 5 = 8$, $\quad 8 =$ settimo numero di Fibonacci,

$5 + 8 = 13$, $\quad 13 =$ ottavo numero di Fibonacci,

$8 + 13 = 21$, $\quad 21 =$ nono numero di Fibonacci,

$13 + 21 = 34$, $\quad 34 =$ decimo numero di Fibonacci,

.....

e la regola si puo' applicare infinite volte, quindi i numeri della sequenza di Fibonacci sono infiniti. Diamo a ciascuno di loro un nome: F_1, F_2, F_3, F_4, ... I primi due numeri della sequenza di Fibonacci sono quindi $F_1 = 0$ e $F_2 = 1$. L'i-esimo ($i = 3, 4, 5, \ldots$) numero della sequenza di Fibonacci, F_i, si ottiene dai precedenti due numeri, F_{i-2} e F_{i-1}, sommandoli:

$F_{i-2} + F_{i-1} = F_i$, $\quad i = 3, 4, 5, \ldots$.

Ad esempio, il terzo numero F_3 ($i = 3$), si ottiene sommando i primi due numeri della sequenza, F_1 e F_2: $F_1 + F_2 = 0 + 1 = 1 = F_3$. Il quarto numero F_4 ($i = 4$), si ottiene sommando F_2 e F_3: $F_2 + F_3 = 1 + 1 = 2 = F_4$. Il quinto numero F_5 ($i = 5$), si ottiene sommando F_3 e F_4: $F_3 + F_4 = 1 + 2 = 3 = F_5$. Il sesto numero F_6 ($i = 6$), si ottiene sommando F_4 e F_5: $F_4 + F_5 = 2 + 3 = 5 = F_6$. E cosi' via ...

Calcolare F_{11}, F_{12} e F_{12}/F_{11}; e poi F_{15}, F_{16} e F_{16}/F_{15}; e poi F_{19}, F_{20} e F_{20}/F_{19}...

Let B be a Hermitian positive definite $n \times n$ matrix, i.e. $B_{ij} = \overline{B_{ji}}$, $i, j = 1, \ldots, n$ ($B = B^H$), and $\mathbf{z}^H B \mathbf{z} > 0 \; \forall \, \mathbf{z} \in \mathbb{C}^n \; \mathbf{z} \neq \mathbf{0}$. Then $B = Q \, \mathrm{diag}\,(\lambda_i, \, i = 1, \ldots, n) Q^H$, where $\lambda_i = \lambda_i(B) > 0$ are the eigevalues of B, $\lambda_i \in \sigma(B)$, and $Q = [\mathbf{q}_1 \, \mathbf{q}_2 \, \cdots \, \mathbf{q}_n]$ is a unitary matrix whose columns $\mathbf{q}_i$ are the eigenvectors of B, $B\mathbf{q}_i = \lambda_i \mathbf{q}_i$ (Q can be chosen real if B is real). Note that B has a Hermitian positive definite square root, that is $B^{\frac{1}{2}} := Q \, \mathrm{diag}\,(\sqrt{\lambda_i}, \, i = 1, \ldots, n) Q^H$.

A well known fact is that the set of real positive numbers $\{\mathbf{u}^H B \mathbf{u} : \; \mathbf{u} \in \mathbb{C}^n \, \|\mathbf{u}\| = 1\}$ coincides with $\overline{\sigma(B)} := [\min_i \lambda_i, \, \max \lambda_i]$, the smallest real interval including the eigenvalues of B. In particular, $\mathbf{q}_i^H B \mathbf{q}_i = \lambda_i$, $i = 1, \ldots, n$. Also observe that $\overline{\sigma(B)}$ includes any real number $(\mathbf{u}^H B^{-1} \mathbf{u})^{-1}$, $\mathbf{u} \in \mathbb{C}^n \; \|\mathbf{u}\| = 1$, in fact $(\mathbf{u}^H B^{-1} \mathbf{u})^{-1} = \mathbf{r}^H B \mathbf{r}$, for $\mathbf{r} = \frac{1}{\sqrt{\mathbf{u}^H B^{-1} \mathbf{u}}} B^{-\frac{1}{2}} \mathbf{u}$, and that such number is smaller than $\mathbf{u}^H B \mathbf{u}$, since, by the Cauchy-Schwarz inequality, $1 = |\mathbf{u}^H \mathbf{u}|^2 = |\mathbf{u}^H B^{\frac{1}{2}} B^{-\frac{1}{2}} \mathbf{u}|^2 \leq \|B^{\frac{1}{2}} \mathbf{u}\|^2 \|B^{-\frac{1}{2}} \mathbf{u}\|^2 = \mathbf{u}^H B \mathbf{u} \, \mathbf{u}^H B^{-1} \mathbf{u}$.

Now let $U = [\mathbf{u}_1 \, \mathbf{u}_2 \, \cdots \, \mathbf{u}_n] \in \mathbb{C}^{n \times n}$ be unitary (for instance U could be the Fourier or the Hartley matrix, or even a Givens or Householder matrix), set $\mathcal{U} = \mathrm{sd}\, U = \{U D U^H : \; D \text{ diagonal}\}$, and consider the following two $n \times n$ Hermitian positive definite matrices of $\mathcal{U}$, associated with B:

$$\mathcal{U}_B = U \, \mathrm{diag}\,(\mathbf{u}_i^H B \mathbf{u}_i) U^H, \qquad (\mathcal{U}_{B^{-1}})^{-1} = U \, \mathrm{diag}\,((\mathbf{u}_i^H B^{-1} \mathbf{u}_i)^{-1}) U^H.$$

Note that $\mathcal{U}_B$ and $\mathcal{U}_{B^{-1}}$ are the projections of the Hermitian positive definite matrices B and B^{-1} into the n-dimensional commutative matrix algebra $\mathcal{U}$, i.e.

$$\|\mathcal{U}_B - B\|_F \leq \|X - B\|_F \; ((X, \mathcal{U}_B - B)_F = 0), \; X \in \mathcal{U}, \quad \|\mathcal{U}_{B^{-1}} - B^{-1}\|_F \leq \|X - B^{-1}\|_F \; ((X, \mathcal{U}_{B^{-1}} - B^{-1})_F = 0), \; X \in \mathcal{U}.$$

The matrices $\mathcal{U}_B$ and $(\mathcal{U}_{B^{-1}})^{-1}$ can be considered approximations of B, since for $U = Q$ they coincide with B, thus, at least in principle, their eigenvalues, $\mathbf{u}_i^H B \mathbf{u}_i$ and $(\mathbf{u}_i^H B^{-1} \mathbf{u}_i)^{-1}$, both included in $\overline{\sigma(B)}$, can be considered approximations of the eigenvalues of B, with $\sigma((\mathcal{U}_{B^{-1}})^{-1})$ shifted on the left with respect to $\sigma(\mathcal{U}_B)$.

Set $\mathbf{r}_i = \frac{1}{\sqrt{\mathbf{u}_i^H B^{-1} \mathbf{u}_i}} B^{-\frac{1}{2}} \mathbf{u}_i$, $\mathbf{h}_i = \frac{1}{\sqrt{\mathbf{u}_i^H B \mathbf{u}_i}} B^{\frac{1}{2}} \mathbf{u}_i$, and let $\min^-$, $\min$, $\max^-$, $\max$ be indeces for which

$$(\mathbf{r}_{\min^-}^H B^{-1} \mathbf{r}_{\min^-})^{-1} \leq (\mathbf{r}_i^H B^{-1} \mathbf{r}_i)^{-1}, \; \mathbf{r}_{\min}^H B \mathbf{r}_{\min} \leq \mathbf{r}_i^H B \mathbf{r}_i, \; (\mathbf{h}_i^H B^{-1} \mathbf{h}_i)^{-1} \leq (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1}, \; \mathbf{h}_i^H B \mathbf{h}_i \leq \mathbf{h}_{\max}^H B \mathbf{h}_{\max},$$

for all i. Note that, by definition of $\mathbf{r}_i$ and $\mathbf{h}_i$, the indeces $\min$ and $\max^-$ are also such that
$(\mathbf{u}_{\min}^H B^{-1} \mathbf{u}_{\min})^{-1} \leq (\mathbf{u}_i^H B^{-1} \mathbf{u}_i)^{-1}$, $\mathbf{u}_i^H B \mathbf{u}_i \leq \mathbf{u}_{\max^-}^H B \mathbf{u}_{\max^-}$, $\forall \, i$.

Define from $\mathcal{U} = \mathrm{sd}\, U$ a space $\mathcal{Z} = \mathrm{sd}\, Z$ such that $\overline{\sigma(\mathcal{U}_B)} \subset \overline{\sigma(\mathcal{Z}_B)}$ or $\overline{\sigma((\mathcal{U}_{B^{-1}})^{-1})} \subset \overline{\sigma((\mathcal{Z}_{B^{-1}})^{-1})}$
Questions: is it always possible? is it possible with the same space $\mathcal{Z}$?.
Try by orthogonalizing the $\mathbf{r}_i$ and the $\mathbf{h}_i$, which form two sets of n linearly independent unit vectors, such that $\mathbf{r}_i^H \mathbf{h}_j = 0 \; i \neq j$, $\mathbf{r}_i^H \mathbf{h}_i = 1 / \sqrt{\mathbf{u}_i^H B \mathbf{u}_i \mathbf{u}_i^H B^{-1} \mathbf{u}_i} \leq 1$.

Results obtained till now:
$\exists \, \mathcal{Z} : \overline{\sigma(\mathcal{U}_B)} \subset \overline{\sigma(\mathcal{Z}_B)}$ whenever $\max = \min = \max^-$ is not verified.
$\exists \, \mathcal{Z} : \overline{\sigma((\mathcal{U}_{B^{-1}})^{-1})} \subset \overline{\sigma((\mathcal{Z}_{B^{-1}})^{-1})}$ whenever $\min = \max^- = \min^-$ is not verified.
$\exists \, \mathcal{Z} : \overline{\sigma(\mathcal{U}_B)} \subset \overline{\sigma(\mathcal{Z}_B)}$ and $\overline{\sigma((\mathcal{U}_{B^{-1}})^{-1})} \subset \overline{\sigma((\mathcal{Z}_{B^{-1}})^{-1})}$ whenever $\max^- \neq \min$.
These conclusions follow from the analysis in the next two pages, which therefore can be skipped at a first reading.

Apply GS to $\{\mathbf{r}_i\}$ starting from $\mathbf{r}_{\min}$: obtain $V = [\mathbf{v}_1 .. \mathbf{r}_{\min} .. \mathbf{v}_n]$ such that $\mathbf{r}_{\min}^H B \mathbf{r}_{\min} = \mathbf{v}_{\min}^H B \mathbf{v}_{\min}$, $\mathbf{r}_i^H B \mathbf{r}_i \leq \mathbf{v}_i^H B \mathbf{v}_i$ $i \neq \min$ (this is conjectured), and thus

$[\leftarrow \mathcal{U}_B]$ $\mathbf{r}_{\min}^H B \mathbf{r}_{\min} = \min_i \lambda_i(\mathcal{V}_B) \leq \lambda_k(\mathcal{U}_B)$ $?\leq? \max_i \lambda_i(\mathcal{V}_B)$,

$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1}]$ $\min_i \lambda_i((\mathcal{V}_{B^{-1}})^{-1}) \leq (\mathbf{v}_{\min}^H B^{-1} \mathbf{v}_{\min})^{-1} \leq \mathbf{r}_{\min}^H B \mathbf{r}_{\min} \leq \lambda_k((\mathcal{U}_{B^{-1}})^{-1})$ $?\leq? \max_i \lambda_i((\mathcal{V}_{B^{-1}})^{-1})$.
(answer to the ? is difficult because the $\mathbf{v}_i$ are generated by GS...)

IF $\max \neq \min$ (so that $\mathbf{h}_{\max}^H \mathbf{r}_{\min} = 0$): apply GS to $\{\mathbf{r}_i\}$ starting from $\mathbf{r}_{\min}$ and then choosing $t, s, .. \neq \max$ (besides $\neq \min$): obtain $V = [\mathbf{v}_1 .. \mathbf{r}_{\min} .. \mathbf{h}_{\max} .. \mathbf{v}_n]$ such that $\mathbf{r}_{\min}^H B \mathbf{r}_{\min} = \mathbf{v}_{\min}^H B \mathbf{v}_{\min}$, $\mathbf{r}_i^H B \mathbf{r}_i \leq \mathbf{v}_i^H B \mathbf{v}_i$ $i \neq \min, \max$ (this is conjectured), $\mathbf{r}_i^H B \mathbf{r}_i \leq \mathbf{h}_{\max}^H B \mathbf{h}_{\max}$, and thus

$[\leftarrow \mathcal{U}_B \rightarrow]$ $\mathbf{r}_{\min}^H B \mathbf{r}_{\min} = \min_i \lambda_i(\tilde{\mathcal{V}}_B) \leq \lambda_k(\mathcal{U}_B) \leq \mathbf{h}_{\max}^H B \mathbf{h}_{\max} \leq \max_i \lambda_i(\tilde{\mathcal{V}}_B)$,

$[\leftarrow (\mathcal{U}_B^{-1})^{-1} \rightarrow!]$ $\min_i \lambda_i((\tilde{\mathcal{V}}_{B^{-1}})^{-1}) \leq (\mathbf{v}_{\min}^H B^{-1} \mathbf{v}_{\min})^{-1} \leq \mathbf{r}_{\min}^H B \mathbf{r}_{\min} \leq \underline{\lambda_k((\mathcal{U}_{B^{-1}})^{-1}) \,!\leq!\, (\mathbf{h}_{\max}^H B^{-1} \mathbf{h}_{\max})^{-1}} \leq$ $\max_i \lambda_i((\tilde{\mathcal{V}}_{B^{-1}})^{-1})$.
(maybe answer to ! is less difficult?...)

ALTERNATIVE: IF $\max^- \neq \min$ (so that $\mathbf{h}_{\max^-}^H \mathbf{r}_{\min} = 0$): apply GS to $\{\mathbf{r}_i\}$ starting from $\mathbf{r}_{\min}$ and then choosing $t, s, .. \neq \max^-$ (besides $\neq \min$): obtain $\tilde{V} = [\mathbf{v}_1 .. \mathbf{r}_{\min} .. \mathbf{h}_{\max^-} .. \mathbf{v}_n]$ such that $\mathbf{r}_{\min}^H B \mathbf{r}_{\min} = \mathbf{v}_{\min}^H B \mathbf{v}_{\min}$, $\mathbf{r}_i^H B \mathbf{r}_i \leq \mathbf{v}_i^H B \mathbf{v}_i$ $i \neq \min, \max$ (this is conjectured), $\mathbf{r}_i^H B \mathbf{r}_i \leq \mathbf{h}_{\max^-}^H B \mathbf{h}_{\max^-}$, and thus

$[\leftarrow \mathcal{U}_B \rightarrow]$ $\mathbf{r}_{\min}^H B \mathbf{r}_{\min} = \min_i \lambda_i(\tilde{\mathcal{V}}_B) \leq \lambda_k(\mathcal{U}_B) \leq \mathbf{h}_{\max^-}^H B \mathbf{h}_{\max^-} \leq \max_i \lambda_i(\tilde{\mathcal{V}}_B)$,

$[\leftarrow (\mathcal{U}_B^{-1})^{-1} \rightarrow]$ $\min_i \lambda_i((\tilde{\mathcal{V}}_{B^{-1}})^{-1}) \leq (\mathbf{v}_{\min}^H B^{-1} \mathbf{v}_{\min})^{-1} \leq \mathbf{r}_{\min}^H B \mathbf{r}_{\min} \leq \lambda_k((\mathcal{U}_{B^{-1}})^{-1}) \leq (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1} \leq$ $\max_i \lambda_i((\tilde{\mathcal{V}}_{B^{-1}})^{-1})$.

Apply GS to $\{\mathbf{h}_i\}$ starting from $\mathbf{h}_{\max^-}$: obtain $W = [\mathbf{w}_1 .. \mathbf{h}_{\max^-} .. \mathbf{w}_n]$ such that $(\mathbf{w}_i^H B^{-1} \mathbf{w}_i)^{-1} \leq (\mathbf{h}_i^H B^{-1} \mathbf{h}_i)^{-1}$ $i \neq \max^-$ (this is conjectured), $(\mathbf{w}_{\max^-}^H B^{-1} \mathbf{w}_{\max^-})^{-1} = (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1}$, and thus

$[(\mathcal{U}_{B^{-1}})^{-1} \rightarrow]$ $\min_i \lambda_i((\mathcal{W}_{B^{-1}})^{-1})? \leq? \lambda_k((\mathcal{U}_{B^{-1}})^{-1}) \leq \max_i \lambda_i((\mathcal{W}_{B^{-1}})^{-1}) = (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1}$,

$[\mathcal{U}_B \rightarrow]$ $\min_i \lambda_i(\mathcal{W}_B)? \leq? \lambda_k(\mathcal{U}_B) \leq (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1} \leq \mathbf{w}_{\max^-}^H B \mathbf{w}_{\max^-} \leq \max_i \lambda_i(\mathcal{W}_B)$.
(answer to the ? is difficult because the $\mathbf{w}_i$ are generated by GS...)

IF $\min^- \neq \max^-$ (so that $\mathbf{r}_{\min^-}^H \mathbf{h}_{\max^-} = 0$): apply GS to $\{\mathbf{h}_i\}$ starting from $\mathbf{h}_{\max^-}$ and then choosing $t, s, .. \neq \min^-$ (besides $\neq \max^-$): obtain $\tilde{W} = [\mathbf{w}_1 .. \mathbf{r}_{\min^-} .. \mathbf{h}_{\max^-} .. \mathbf{w}_n]$ such that $(\mathbf{w}_i^H B^{-1} \mathbf{w}_i)^{-1} \leq (\mathbf{h}_i^H B^{-1} \mathbf{h}_i)^{-1}$ $i \neq \max^-, \min^-$ (this is conjectured), $(\mathbf{w}_{\max^-}^H B^{-1} \mathbf{w}_{\max^-})^{-1} = (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1}$, $(\mathbf{r}_{\min^-}^H B^{-1} \mathbf{r}_{\min^-})^{-1} \leq$ $(\mathbf{h}_i^H B^{-1} \mathbf{h}_i)^{-1}$, and thus

$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1} \rightarrow] \min_i \lambda_i((\tilde{\mathcal{W}}_{B^{-1}})^{-1}) \leq (\mathbf{r}_{\min^-}^T B^{-1} \mathbf{r}_{\min^-})^{-1} \leq \lambda_k((\mathcal{U}_{B^{-1}})^{-1}) \leq \max_i \lambda_i((\tilde{\mathcal{W}}_{B^{-1}})^{-1}) = (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1}$.

$[! \leftarrow \mathcal{U}_B \rightarrow] \min_i \lambda_i(\tilde{\mathcal{W}}_B) \leq \underline{\mathbf{r}_{\min^-}^H B \mathbf{r}_{\min^-} ! \leq! \lambda_k(\mathcal{U}_B)} \leq (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1} \leq \mathbf{w}_{\max^-}^H B \mathbf{w}_{\max^-} \leq \max_i \lambda_i(\tilde{\mathcal{W}}_B)$
(maybe answer to ! is less difficult?...)

ALTERNATIVE: IF $\min \neq \max^-$ (so that $\mathbf{r}_{\min}^H \mathbf{h}_{\max^-} = 0$): apply GS to $\{\mathbf{h}_i\}$ starting from $\mathbf{h}_{\max^-}$ and then choosing $t, s, .. \neq \min$ (besides $\neq \max^-$): obtain $\tilde{W} = [\mathbf{w}_1 .. \mathbf{r}_{\min} .. \mathbf{h}_{\max^-} .. \mathbf{w}_n]$ such that $(\mathbf{w}_i^H B^{-1} \mathbf{w}_i)^{-1} \leq$ $(\mathbf{h}_i^H B^{-1} \mathbf{h}_i)^{-1}$ $i \neq \max^-, \min$ (this is conjectured), $(\mathbf{w}_{\max^-}^H B^{-1} \mathbf{w}_{\max^-})^{-1} = (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1}$, $(\mathbf{r}_{\min}^H B^{-1} \mathbf{r}_{\min})^{-1} \leq$ $(\mathbf{h}_i^H B^{-1} \mathbf{h}_i)^{-1}$, and thus

$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1} \rightarrow] \min_i \lambda_i((\tilde{\mathcal{W}}_{B^{-1}})^{-1}) \leq (\mathbf{r}_{\min}^T B^{-1} \mathbf{r}_{\min})^{-1} \leq \lambda_k((\mathcal{U}_{B^{-1}})^{-1}) \leq \max_i \lambda_i((\tilde{\mathcal{W}}_{B^{-1}})^{-1}) = (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1}$.

$[\leftarrow \mathcal{U}_B \rightarrow] \min_i \lambda_i(\tilde{\mathcal{W}}_B) \leq \mathbf{r}_{\min}^H B \mathbf{r}_{\min} \leq \lambda_k(\mathcal{U}_B) \leq (\mathbf{h}_{\max^-}^H B^{-1} \mathbf{h}_{\max^-})^{-1} \leq \mathbf{w}_{\max^-}^H B \mathbf{w}_{\max^-} \leq \max_i \lambda_i(\tilde{\mathcal{W}}_B)$

Until the line we have $(\mathbf{z}_i^H B^{-1}\mathbf{z}_i)^{-1} \le (\mathbf{h}_i^H B^{-1}\mathbf{h}_i)^{-1}$, with $=$ if $i = \max$ (to be verified)
$[\mathcal{U}_B \rightarrow]$: GS to $\mathbf{h}_{\max}$, $\mathbf{h}_i$ $i \neq \max$, yields orthonormal $\mathbf{z}_{\max} = \mathbf{h}_{\max}$, $\mathbf{z}_i$ $i \neq \max$ such that

$$\min_i \lambda_i(\mathcal{Z}_B)? \le? \ \lambda_k(\mathcal{U}_B) \le \mathbf{h}_{\max}^H B \mathbf{h}_{\max} \le \max_i \lambda_i(\mathcal{Z}_B)$$

where the last inequality can be strict (or not?).
$[(\mathcal{U}_{B^{-1}})^{-1} \rightarrow!]$ $\min_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1})? \le? \ \underline{\lambda_k((\mathcal{U}_{B^{-1}})^{-1})! \le!(\mathbf{h}_{\max}^H B^{-1}\mathbf{h}_{\max})^{-1}} \le \max_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1})$

$[\leftarrow \mathcal{U}_B \rightarrow]$: If $\max \neq \min$ (so that $\mathbf{h}_{\max}^H \mathbf{r}_{\min} = 0$): GS to $\mathbf{h}_{\max}$, $\mathbf{r}_{\min}$, $\mathbf{h}_i$ $i \neq \max, \min$, yields orthonormal $\mathbf{z}_{\max} = \mathbf{h}_{\max}$, $\mathbf{z}_{\min} = \mathbf{r}_{\min}$, $\mathbf{z}_i$ $i \neq \max, \min$ such that

$$\min_i \lambda_i(\mathcal{Z}_B) \le \mathbf{r}_{\min}^H B \mathbf{r}_{\min} \le \lambda_k(\mathcal{U}_B) \le \mathbf{h}_{\max}^H B \mathbf{h}_{\max} \le \max_i \lambda_i(\mathcal{Z}_B)$$

where the first and the last inequality can be strict (or not?).
$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1} \rightarrow!]$ $\min_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1}) \le (\mathbf{r}_{\min}^H B^{-1}\mathbf{r}_{\min})^{-1} \le \underline{\lambda_k((\mathcal{U}_{B^{-1}})^{-1})! \le!(\mathbf{h}_{\max}^H B^{-1}\mathbf{h}_{\max})^{-1}} \le \max_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1})$

$[! \leftarrow \mathcal{U}_B \rightarrow]$: If $\max \neq \min^-$ (so that $\mathbf{h}_{\max}^H \mathbf{r}_{\min^-} = 0$): GS to $\mathbf{h}_{\max}$, $\mathbf{r}_{\min^-}$, $\mathbf{h}_i$ $i \neq \max, \min^-$, yields orthonormal $\mathbf{z}_{\max} = \mathbf{h}_{\max}$, $\mathbf{z}_{\min^-} = \mathbf{r}_{\min^-}$, $\mathbf{z}_i$ $i \neq \max, \min^-$ such that

$$\min_i \lambda_i(\mathcal{Z}_B) \le \underline{\mathbf{r}_{\min^-}^H B \mathbf{r}_{\min^-}! \le!\lambda_k(\mathcal{U}_B)} \le \mathbf{h}_{\max}^H B \mathbf{h}_{\max} \le \max_i \lambda_i(\mathcal{Z}_B)$$

where the first and the last inequality can be strict (or not?).
$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1} \rightarrow!]$ $\min_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1}) \le (\mathbf{r}_{\min^-}^H B^{-1}\mathbf{r}_{\min^-})^{-1} \le \underline{\lambda_k((\mathcal{U}_{B^{-1}})^{-1})! \le!(\mathbf{h}_{\max}^H B^{-1}\mathbf{h}_{\max})^{-1}} \le \max_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1})$

Question: can the inequalities $(\mathbf{h}_{\max}^H B^{-1}\mathbf{h}_{\max})^{-1} < (\mathbf{h}_{\max^-}^H B^{-1}\mathbf{h}_{\max^-})^{-1}$, $\mathbf{r}_{\min}^H B \mathbf{r}_{\min} < \mathbf{r}_{\min^-}^H B \mathbf{r}_{\min^-}$ hold?

Until the line we have $\mathbf{r}_i^H B \mathbf{r}_i \le \mathbf{z}_i^H B \mathbf{z}_i$, with $=$ if $i = \min^-$ (to be verified)
$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1}]$: GS to $\mathbf{r}_{\min^-}$, $\mathbf{r}_i$ $i \neq \min^-$, yields orthonormal $\mathbf{z}_{\min^-} = \mathbf{r}_{\min^-}$, $\mathbf{z}_i$ $i \neq \min^-$ such that

$$\min_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1}) \le (\mathbf{r}_{\min^-}^H B^{-1}\mathbf{r}_{\min^-})^{-1} \le \lambda_k((\mathcal{U}_{B^{-1}})^{-1}) \ ? \le? \max_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1})$$

where the first inequality can be strict (or not?).
$[! \leftarrow \mathcal{U}_B]$ $\min \lambda_i(\mathcal{Z}_B) \le \underline{\mathbf{r}_{\min^-}^H B \mathbf{r}_{\min^-}! \le!\lambda_k(\mathcal{U}_B)}? \le? \max_i \lambda_i(\mathcal{Z}_B)$

$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1} \rightarrow]$: If $\max^- \neq \min^-$ (so that $\mathbf{r}_{\min^-}^H \mathbf{h}_{\max^-} = 0$): GS to $\mathbf{r}_{\min^-}$, $\mathbf{h}_{\max^-}$, $\mathbf{r}_i$ $i \neq \min^-, \max^-$, yields orthonormal $\mathbf{z}_{\min^-} = \mathbf{r}_{\min^-}$, $\mathbf{z}_{\max^-} = \mathbf{h}_{\max^-}$, $\mathbf{z}_i$ $i \neq \min^-, \max^-$ such that

$$\min_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1}) \le (\mathbf{r}_{\min^-}^H B^{-1}\mathbf{r}_{\min^-})^{-1} \le \lambda_k((\mathcal{U}_{B^{-1}})^{-1}) \le (\mathbf{h}_{\max^-}^H B^{-1}\mathbf{h}_{\max^-})^{-1} \le \max_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1})$$

where the first and the last inequality can be strict (or not?).
$[! \leftarrow \mathcal{U}_B \rightarrow]$ $\min \lambda_i(\mathcal{Z}_B) \le \underline{\mathbf{r}_{\min^-}^H B \mathbf{r}_{\min^-}! \le!\lambda_k(\mathcal{U}_B)} \le \mathbf{h}_{\max^-}^H B \mathbf{h}_{\max^-} \le \max_i \lambda_i(\mathcal{Z}_B)$

$[\leftarrow (\mathcal{U}_{B^{-1}})^{-1} \rightarrow!]$: If $\max \neq \min^-$ (so that $\mathbf{r}_{\min^-}^H \mathbf{h}_{\max} = 0$): GS to $\mathbf{r}_{\min^-}$, $\mathbf{h}_{\max}$, $\mathbf{r}_i$ $i \neq \min^-, \max$, yields orthonormal $\mathbf{z}_{\min^-} = \mathbf{r}_{\min^-}$, $\mathbf{z}_{\max} = \mathbf{h}_{\max}$, $\mathbf{z}_i$ $i \neq \min^-, \max$ such that

$$\min_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1}) \le (\mathbf{r}_{\min^-}^H B^{-1}\mathbf{r}_{\min^-})^{-1} \le \underline{\lambda_k((\mathcal{U}_{B^{-1}})^{-1})! \le!(\mathbf{h}_{\max}^H B^{-1}\mathbf{h}_{\max})^{-1}} \le \max_i \lambda_i((\mathcal{Z}_{B^{-1}})^{-1})$$

where the first and the last inequality can be strict (or not?).
$[! \leftarrow \mathcal{U}_B \rightarrow]$ $\min \lambda_i(\mathcal{Z}_B) \le \underline{\mathbf{r}_{\min^-}^H B \mathbf{r}_{\min^-}! \le!\lambda_k(\mathcal{U}_B)} \le \mathbf{h}_{\max}^H B \mathbf{h}_{\max} \le \max_i \lambda_i(\mathcal{Z}_B)$

If $\mathbf{r} = \mathbf{r}_C = \frac{1}{\sqrt{\mathbf{u}^T B^{-1}\mathbf{u}}} B^{-\frac{1}{2}} C\mathbf{u}$, with C chosen such that $\|Q^T C\mathbf{u}\|^2 = \sum_r [Q^T C\mathbf{u}]_r^2 = 1$ and $\sum_r \frac{1}{\lambda_r}[Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T C^T B^{-1} C\mathbf{u} = \mathbf{u}^T B^{-1}\mathbf{u}$, then

$$\mathbf{r}^T\mathbf{r} = \frac{\mathbf{u}^T C^T B^{-1} C\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}} = 1, \quad (\frac{\mathbf{u}^T C^T B^{-2} C\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1} = (\mathbf{r}^T B^{-1}\mathbf{r})^{-1} \leq \mathbf{r}^T B\mathbf{r} = (\mathbf{u}^T B^{-1}\mathbf{u})^{-1}.$$

If $\mathbf{h} = \mathbf{h}_C = \frac{1}{\sqrt{\mathbf{u}^T B\mathbf{u}}} B^{\frac{1}{2}} C\mathbf{u}$, with C chosen such that $\|Q^T C\mathbf{u}\|^2 = \sum_r [Q^T C\mathbf{u}]_r^2 = 1$ and $\sum_r \lambda_r [Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T C^T B C\mathbf{u} = \mathbf{u}^T B\mathbf{u}$, then

$$\mathbf{h}^T\mathbf{h} = \frac{\mathbf{u}^T C^T B C\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} = 1, \quad \mathbf{u}^T B\mathbf{u} = (\mathbf{h}^T B^{-1}\mathbf{h})^{-1} \leq \mathbf{h}^T B\mathbf{h} = \frac{\mathbf{u}^T C^T B^2 C\mathbf{u}}{\mathbf{u}^T B\mathbf{u}}.$$

For $C = I$ the assumptions are verified, and in the above inequalities one recognizes one step of the power method, applied to B^{-1} ($\min_i \lambda_i \leftarrow (\frac{\mathbf{u}^T B^{-2}\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1} \leftarrow (\mathbf{u}^T B^{-1}\mathbf{u})^{-1}$) and to B ($\mathbf{u}^T B\mathbf{u} \rightarrow \frac{\mathbf{u}^T B^2\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} \rightarrow \max_i \lambda_i$).

In order to try to improve the power method applied to B^{-1}, choose $C = C^-$ such that $(\mathbf{r}^T B^{-1}\mathbf{r})^{-1}$ is minimum or, equivalently, such that $\mathbf{u}^T C^T B^{-2} C\mathbf{u} = \sum_r \frac{1}{\lambda_r^2}[Q^T C\mathbf{u}]_r^2$ is maximum. Analogously, in order to improve the power method applied to B, choose $C = C^+$ such that $\mathbf{h}^T B\mathbf{h}$ is maximum or, equivalently, such that $\mathbf{u}^T C^T B^2 C\mathbf{u} = \sum_r \lambda_r^2 [Q^T C\mathbf{u}]_r^2$ is maximum.

Lemma (improve power to B: $\mathbf{u}^T B\mathbf{u} \rightarrow \frac{\mathbf{u}^T B^2\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} \rightarrow \frac{\mathbf{u}^T C_o^T B^2 C_o\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} \rightarrow \max_i \lambda_i$)
Let m and M be indices such that $\lambda_m \leq \lambda_i \leq \lambda_M$, $\forall\, i$. If $\sum_r x_r^2 = 1$ and $\sum_r \lambda_r x_r^2 = \mathbf{u}^T B\mathbf{u}$, then $\sum_r \lambda_r^2 x_r^2 + \sum_{r\neq M,m} x_r^2(\lambda_M - \lambda_r)(\lambda_r - \lambda_m) = \mathbf{u}^T B\mathbf{u}(\lambda_M + \lambda_m) - \lambda_M\lambda_m = \lambda_M^2 \frac{\mathbf{u}^T B\mathbf{u} - \lambda_m}{\lambda_M - \lambda_m} + \lambda_m^2 \frac{\lambda_M - \mathbf{u}^T B\mathbf{u}}{\lambda_M - \lambda_m}$, where the last equality holds only if $\lambda_M \neq \lambda_m$. As a consequence, if $\lambda_m < \lambda_M$, then the conditions on the x_k

$$\sum_r x_r^2 = 1, \quad \sum_r \lambda_r x_r^2 = \mathbf{u}^T B\mathbf{u}, \quad \sum_r \lambda_r^2 x_r^2 \text{ is maximum} \qquad (\#)$$

are satisfied for $x_M^2 = \frac{\mathbf{u}^T B\mathbf{u} - \lambda_m}{\lambda_M - \lambda_m}$, $x_r = 0$ $r \neq M, m$, $x_m^2 = \frac{\lambda_M - \mathbf{u}^T B\mathbf{u}}{\lambda_M - \lambda_m}$, and are not satisfied for other values of x_r whenever λ_m and λ_M are simple.
Let $C_o = C_{ott}$ be a matrix such that $\sum_r [Q^T C_o\mathbf{u}]_r^2 = 1$, $\sum_r \lambda_r [Q^T C_o\mathbf{u}]_r^2 = \mathbf{u}^T B\mathbf{u}$, and

$$\sum_r \lambda_r^2 [Q^T C_o\mathbf{u}]_r^2 = \mathbf{u}^T C_o^T B^2 C_o\mathbf{u} \geq \sum_r \lambda_r^2 [Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T C^T B^2 C\mathbf{u}$$

for all C such that $\sum_r [Q^T C\mathbf{u}]_r^2 = 1$, $\sum_r \lambda_r [Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T B\mathbf{u}$, i.e., by the above arguments,

$$C_o\mathbf{u} = x_m\mathbf{q}_m + x_M\mathbf{q}_M, \quad x_m^2 = \frac{\lambda_M - \mathbf{u}^T B\mathbf{u}}{\lambda_M - \lambda_m}, \quad x_M^2 = \frac{\mathbf{u}^T B\mathbf{u} - \lambda_m}{\lambda_M - \lambda_m}.$$

Then, for any matrix C such that $\sum_r [Q^T C\mathbf{u}]_r^2 = 1$, $\sum_r \lambda_r [Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T B\mathbf{u}$, we have

$$\mathbf{u}^T B\mathbf{u} = (\mathbf{h}_C^T B^{-1}\mathbf{h}_C)^{-1} \leq \mathbf{h}_C^T B\mathbf{h}_C = \frac{\mathbf{u}^T C^T B^2 C\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} \leq \mathbf{h}_{C_o}^T B\mathbf{h}_{C_o} = \frac{\mathbf{u}^T C_o^T B^2 C_o\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} = \lambda_m + \lambda_M - \frac{\lambda_m\lambda_M}{\mathbf{u}^T B\mathbf{u}} \leq \lambda_M$$

(used properties: $0 < \lambda_m < \lambda_M$, $\mathbf{u}^T B\mathbf{u}, \lambda_r \in (\lambda_m, \lambda_M)$ $r \neq m, M$; satisfied by other entities...?).

C_o is well defined (for ex as a Householder matrix), but we do not know λ_m, $\mathbf{q}_m$, λ_M, $\mathbf{q}_M$.

proof. The first equality in case $\lambda_m = \lambda_M$ is left to the reader. For $\lambda_m < \lambda_M$, by using the first two conditions in (#), obtain (via Kramer) expressions for x_M^2 and x_m^2, and replace them in $\sum_r \lambda_r^2 x_r^2$:

$\sum_r \lambda_r^2 x_r^2 = \lambda_M^2 \frac{1}{\lambda_M - \lambda_m}[(\mathbf{u}^T B\mathbf{u} - \sum_{r\neq M,m} \lambda_r x_r^2) - (1 - \sum_{r\neq M,m} x_r^2)\lambda_m]$
$+ \sum_{r\neq M,m} \lambda_r^2 x_r^2 + \lambda_m^2 \frac{1}{\lambda_M - \lambda_m}[(1 - \sum_{r\neq M,m} x_r^2)\lambda_M - (\mathbf{u}^T B\mathbf{u} - \sum_{r\neq M,m} \lambda_r x_r^2)]$
$= \lambda_M^2 \frac{\mathbf{u}^T B\mathbf{u} - \lambda_m}{\lambda_M - \lambda_m} + \lambda_m^2 \frac{\lambda_M - \mathbf{u}^T B\mathbf{u}}{\lambda_M - \lambda_m} + \lambda_M^2 \frac{1}{\lambda_M - \lambda_m} \sum_{r\neq M,m} x_r^2(\lambda_m - \lambda_r)$
$+ \sum_{r\neq M,m} \lambda_r^2 x_r^2 + \lambda_m^2 \frac{1}{\lambda_M - \lambda_m} \sum_{r\neq M,m} x_r^2(\lambda_r - \lambda_M)$
$= \lambda_M^2 \frac{\mathbf{u}^T B\mathbf{u} - \lambda_m}{\lambda_M - \lambda_m} + \lambda_m^2 \frac{\lambda_M - \mathbf{u}^T B\mathbf{u}}{\lambda_M - \lambda_m} + \sum_{r\neq M,m} x_r^2[...]$ where
$[...] = \frac{\lambda_M^2(\lambda_m - \lambda_r)}{\lambda_M - \lambda_m} + \lambda_r^2 + \frac{\lambda_m^2(\lambda_r - \lambda_M)}{\lambda_M - \lambda_m}$
$= \frac{\lambda_M^2(\lambda_m - \lambda_r) + \lambda_r^2(\lambda_M - \lambda_r) + \lambda_r^2(\lambda_r - \lambda_m) + \lambda_m^2(\lambda_r - \lambda_M)}{\lambda_M - \lambda_m}$
$= \frac{(\lambda_m - \lambda_r)(\lambda_M^2 - \lambda_r^2) + (\lambda_M - \lambda_r)(\lambda_r^2 - \lambda_m^2)}{\lambda_M - \lambda_m}$
$= -(\lambda_M - \lambda_r)(\lambda_r - \lambda_m).$

Lemma (improve power to B^{-1}: $\min_i \lambda_i \leftarrow (\frac{\mathbf{u}^T C_o^T B^{-2} C_o \mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1} \leftarrow (\frac{\mathbf{u}^T B^{-2}\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1} \leftarrow (\mathbf{u}^T B^{-1}\mathbf{u})^{-1}$)
Let m and M be indices such that $\lambda_m \leq \lambda_i \leq \lambda_M$, $\forall\, i$. If $\sum_r x_r^2 = 1$ and $\sum_r \frac{1}{\lambda_r} x_r^2 = \mathbf{u}^T B^{-1}\mathbf{u}$, then $\sum_r \frac{1}{\lambda_r^2} x_r^2 + \sum_{r\neq M,m} x_r^2 (\frac{1}{\lambda_M} - \frac{1}{\lambda_r})(\frac{1}{\lambda_r} - \frac{1}{\lambda_m}) = \mathbf{u}^T B^{-1}\mathbf{u}(\frac{1}{\lambda_M} + \frac{1}{\lambda_m}) - \frac{1}{\lambda_M \lambda_m} = \frac{1}{\lambda_M^2} \frac{\mathbf{u}^T B^{-1}\mathbf{u} - \frac{1}{\lambda_m}}{\frac{1}{\lambda_M} - \frac{1}{\lambda_m}} + \frac{1}{\lambda_m^2} \frac{\frac{1}{\lambda_M} - \mathbf{u}^T B^{-1}\mathbf{u}}{\frac{1}{\lambda_M} - \frac{1}{\lambda_m}}$,
where the last equality holds only if $\lambda_M \neq \lambda_m$. As a consequence, if $\lambda_m < \lambda_M$, then the conditions on the x_k

$$\sum_r x_r^2 = 1, \quad \sum_r \frac{1}{\lambda_r} x_r^2 = \mathbf{u}^T B^{-1}\mathbf{u}, \quad \sum_r \frac{1}{\lambda_r^2} x_r^2 \text{ is maximum} \qquad (\#)$$

are satisfied for $x_M^2 = \frac{\mathbf{u}^T B^{-1}\mathbf{u} - \frac{1}{\lambda_m}}{\frac{1}{\lambda_M} - \frac{1}{\lambda_m}}$, $x_r = 0$ $r \neq M, m$, $x_m^2 = \frac{\frac{1}{\lambda_M} - \mathbf{u}^T B^{-1}\mathbf{u}}{\frac{1}{\lambda_M} - \frac{1}{\lambda_m}}$, and are not satisfied for other values of x_r whenever λ_m and λ_M are simple.
Let $C_o = C_{ott}$ be a matrix such that $\sum_r [Q^T C_o \mathbf{u}]_r^2 = 1$, $\sum_r \frac{1}{\lambda_r} [Q^T C_o \mathbf{u}]_r^2 = \mathbf{u}^T B^{-1}\mathbf{u}$, and

$$\sum_r \frac{1}{\lambda_r^2} [Q^T C_o \mathbf{u}]_r^2 = \mathbf{u}^T C_o^T B^{-2} C_o \mathbf{u} \geq \sum_r \frac{1}{\lambda_r^2} [Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T C^T B^{-2} C\mathbf{u}$$

for all C such that $\sum_r [Q^T C\mathbf{u}]_r^2 = 1$, $\sum_r \frac{1}{\lambda_r} [Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T B^{-1}\mathbf{u}$, i.e.

$$C_o\mathbf{u} = x_m \mathbf{q}_m + x_M \mathbf{q}_M, \quad x_m^2 = \frac{\frac{1}{\lambda_M} - \mathbf{u}^T B^{-1}\mathbf{u}}{\frac{1}{\lambda_M} - \frac{1}{\lambda_m}}, \quad x_M^2 = \frac{\mathbf{u}^T B^{-1}\mathbf{u} - \frac{1}{\lambda_m}}{\frac{1}{\lambda_M} - \frac{1}{\lambda_m}}.$$

Then, for any matrix C such that $\sum_r [Q^T C\mathbf{u}]_r^2 = 1$, $\sum_r \frac{1}{\lambda_r} [Q^T C\mathbf{u}]_r^2 = \mathbf{u}^T B^{-1}\mathbf{u}$, we have

$$\lambda_m \leq \frac{\lambda_M \lambda_m}{\lambda_M + \lambda_m - (\mathbf{u}^T B^{-1}\mathbf{u})^{-1}} = \frac{1}{\frac{1}{\lambda_m} + \frac{1}{\lambda_M} - \frac{1}{\lambda_m \lambda_M \mathbf{u}^T B^{-1}\mathbf{u}}} = (\frac{\mathbf{u}^T C_o^T B^{-2} C_o \mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1} = (\mathbf{r}_{C_o}^T B^{-1} \mathbf{r}_{C_o})^{-1}$$

$$\leq (\frac{\mathbf{u}^T C^T B^{-2} C\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1} = (\mathbf{r}_C^T B^{-1} \mathbf{r}_C)^{-1} \leq \mathbf{r}_C^T B \mathbf{r}_C = (\mathbf{u}^T B^{-1}\mathbf{u})^{-1}$$

(used properties: $0 < \lambda_m < \lambda_M$, $\mathbf{u}^T B^{-1}\mathbf{u}$, $\frac{1}{\lambda_r} \in (\frac{1}{\lambda_M}, \frac{1}{\lambda_m})$ $r \neq m, M$; satisfied by other entities...?).

C_o is well defined (for ex as a Householder matrix), but we do not know λ_m, $\mathbf{q}_m$, λ_M, $\mathbf{q}_M$.

Our aim is to improve the approximation $\mathbf{u}^T B\mathbf{u}$ of λ_M with something better than $\frac{\mathbf{u}^T B^2\mathbf{u}}{\mathbf{u}^T B\mathbf{u}}$. The above Lemma states that

$$\mathbf{h}_{C_o}^T B\mathbf{h}_{C_o} = \frac{\mathbf{u}^T C_o^T B^2 C_o\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} = \lambda_m + \lambda_M - \frac{\lambda_m\lambda_M}{\mathbf{u}^T B\mathbf{u}}, \quad C_o\mathbf{u} = \sqrt{\frac{\lambda_M - \mathbf{u}^T B\mathbf{u}}{\lambda_M - \lambda_m}}\mathbf{q}_m + \sqrt{\frac{\mathbf{u}^T B\mathbf{u} - \lambda_m}{\lambda_M - \lambda_m}}\mathbf{q}_M,$$

is certainly better than $\mathbf{h}_I^T B\mathbf{h}_I = \frac{\mathbf{u}^T B^2\mathbf{u}}{\mathbf{u}^T B\mathbf{u}}$, but of course such number is not computable. However we can replace the unknown λ_m, $\mathbf{q}_m$, λ_M, $\mathbf{q}_M$, present in the definition of such number, with some approximations of them computable from what is available, and in particular from $\mathbf{u}^T B\mathbf{u}$ and $\frac{\mathbf{u}^T B^2\mathbf{u}}{\mathbf{u}^T B\mathbf{u}}$.

Assuming to have $\mathcal{U}_B = U\,\mathrm{diag}\,(\mathbf{u}_i^H B\mathbf{u}_i,\, i = 1,\dots,n)U^H$, we can think $\mathbf{u}^T B\mathbf{u}$ as one of the eigenvalues of $\mathcal{U}_B$, i.e. we can think $\mathbf{u} = \mathbf{u}_i$ (for instance $\mathbf{u} = \mathbf{u}_{\max^-}$, recalling the definition of $\max^-$, $\mathbf{u}_i^T B\mathbf{u}_i \le \mathbf{u}_{\max^-}^T B\mathbf{u}_{\max^-}\ \forall\, i$).

Assume that we can construct a unitary matrix $\tilde{Q} = [\tilde{\mathbf{q}}_1\ \tilde{\mathbf{q}}_2\ \cdots\ \tilde{\mathbf{q}}_n]$ such that $\tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner = \min_i \tilde{\mathbf{q}}_i^T B\tilde{\mathbf{q}}_i < \mathbf{u}^T B\mathbf{u} < \max_i \tilde{\mathbf{q}}_i^T B\tilde{\mathbf{q}}_i = \tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner$. (For example, if $\max \neq \min$, then obtain $\tilde{Q}$ applying GS to $\mathbf{r}_{\min}$, $\mathbf{r}_i$ $i \neq \min, \max$, and adding $\mathbf{h}_{\max}$, or applying GS to $\mathbf{h}_{\max}$, $\mathbf{h}_i$ $i \neq \min, \max$, and adding $\mathbf{r}_{\min}$.) Then the idea is to consider $\tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner$, $\tilde{\mathbf{q}}_\llcorner$, $\tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner$, $\tilde{\mathbf{q}}_\lrcorner$, as approximations of λ_m, $\mathbf{q}_m$, λ_M, $\mathbf{q}_M$, respectively, and define $\tilde{C}_o$ so that

$$\tilde{C}_o\mathbf{u} = \sqrt{\frac{\tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner - \mathbf{u}^T B\mathbf{u}}{\tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner - \tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner}}\tilde{\mathbf{q}}_\llcorner + \sqrt{\frac{\mathbf{u}^T B\mathbf{u} - \tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner}{\tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner - \tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner}}\tilde{\mathbf{q}}_\lrcorner,$$

Observe that $\|\tilde{Q}^T\tilde{C}_o\mathbf{u}\|^2 = \mathbf{u}^T\tilde{C}_o^T\tilde{Q}\tilde{Q}^T\tilde{C}_o\mathbf{u} = 1$, $\mathbf{u}^T\tilde{C}_o^T\tilde{\mathcal{Q}}_B\tilde{C}_o\mathbf{u} = \mathbf{u}^T B\mathbf{u}$, and

$$\mathbf{u}^T\tilde{C}_o^T\tilde{\mathcal{Q}}_B^2\tilde{C}_o\mathbf{u} = (\tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner + \tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner)\mathbf{u}^T B\mathbf{u} - \tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner\tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner$$

$$\ge \mathbf{u}^T\tilde{C}^T\tilde{\mathcal{Q}}_B^2\tilde{C}\mathbf{u} = -\sum_{r\neq\llcorner,\lrcorner}[\tilde{Q}^T\tilde{C}\mathbf{u}]_r^2(\tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner - \tilde{\mathbf{q}}_r^T B\tilde{\mathbf{q}}_r)(\tilde{\mathbf{q}}_r^T B\tilde{\mathbf{q}}_r - \tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner)$$

$$+\ \ (\tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner + \tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner)\mathbf{u}^T B\mathbf{u} - \tilde{\mathbf{q}}_\lrcorner^T B\tilde{\mathbf{q}}_\lrcorner\tilde{\mathbf{q}}_\llcorner^T B\tilde{\mathbf{q}}_\llcorner$$

for all $\tilde{C}$ such that $\mathbf{u}^T\tilde{C}^T\tilde{Q}\tilde{Q}^T\tilde{C}\mathbf{u} = 1$, $\mathbf{u}^T\tilde{C}^T\tilde{\mathcal{Q}}_B\tilde{C}\mathbf{u} = \mathbf{u}^T B\mathbf{u}$.

Problem: $\mathbf{u}^T\tilde{C}_o^T QQ^T\tilde{C}_o\mathbf{u} = 1$ ok, but $\mathbf{u}^T\tilde{C}_o^T B\tilde{C}_o\mathbf{u} = \mathbf{u}^T B\mathbf{u}$ not ok, $\mathbf{u}^T\tilde{C}_o^T B^2\tilde{C}_o\mathbf{u}$ not maximum; so what inequalities the new actors satisfy in place of

$$\mathbf{u}^T B\mathbf{u} = (\mathbf{h}_C^T B^{-1}\mathbf{h}_C)^{-1} \le \mathbf{h}_C^T B\mathbf{h}_C = \frac{\mathbf{u}^T C^T B^2 C\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} \le \mathbf{h}_{\tilde{C}_o}^T B\mathbf{h}_{\tilde{C}_o} = \frac{\mathbf{u}^T\tilde{C}_o^T B^2\tilde{C}_o\mathbf{u}}{\mathbf{u}^T B\mathbf{u}} = \lambda_m + \lambda_M - \frac{\lambda_m\lambda_M}{\mathbf{u}^T B\mathbf{u}} \le \lambda_M$$

for all C such that $\mathbf{u}^T C^T QQ^T C\mathbf{u} = 1$, $\mathbf{u}^T C^T BC\mathbf{u} = \mathbf{u}^T B\mathbf{u}$?

Our aim is to improve the approximation $(\mathbf{u}^T B^{-1}\mathbf{u})^{-1}$ of λ_m with something better than $(\frac{\mathbf{u}^T B^{-2}\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1}$. The above Lemma states that

$$\frac{\lambda_M\lambda_m}{\lambda_M+\lambda_m-(\mathbf{u}^T B^{-1}\mathbf{u})^{-1}}=\frac{1}{\frac{1}{\lambda_m}+\frac{1}{\lambda_M}-\frac{1}{\lambda_m\lambda_M\mathbf{u}^T B^{-1}\mathbf{u}}}=(\frac{\mathbf{u}^T C_o^T B^{-2}C_o\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1}$$

$$=(\mathbf{r}_{C_o}^T B^{-1}\mathbf{r}_{C_o})^{-1},\ \ C_o\mathbf{u}=\sqrt{\frac{\frac{1}{\lambda_M}-\mathbf{u}^T B^{-1}\mathbf{u}}{\frac{1}{\lambda_M}-\frac{1}{\lambda_m}}}\mathbf{q}_m+\sqrt{\frac{\mathbf{u}^T B^{-1}\mathbf{u}-\frac{1}{\lambda_m}}{\frac{1}{\lambda_M}-\frac{1}{\lambda_m}}}\mathbf{q}_M,$$

is certainly better than $(\mathbf{r}_I^T B^{-1}\mathbf{r}_I)^{-1}=(\frac{\mathbf{u}^T B^{-2}\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1}$, but of course such number is not computable. However we can replace the unknown λ_m, $\mathbf{q}_m$, λ_M, $\mathbf{q}_M$, present in the definition of such number, with some approximations of them computable from what is available, and in particular from $(\mathbf{u}^T B^{-1}\mathbf{u})^{-1}$ and $(\frac{\mathbf{u}^T B^{-2}\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1}$.

Assuming to have $(\mathcal{U}_{B^{-1}})^{-1}=U\,\mathrm{diag}\,((\mathbf{u}_i^H B^{-1}\mathbf{u}_i)^{-1},\ i=1,\dots,n)U^H$, we can think $(\mathbf{u}^T B^{-1}\mathbf{u})^{-1}$ as one of the eigenvalues of $(\mathcal{U}_{B^{-1}})^{-1}$, i.e. we can think $\mathbf{u}=\mathbf{u}_i$ (for instance $\mathbf{u}=\mathbf{u}_{\min}$, recalling the definition of min, $(\mathbf{u}_{\min}^T B^{-1}\mathbf{u}_{\min})^{-1}\le(\mathbf{u}_i^T B^{-1}\mathbf{u}_i)^{-1}\ \forall\, i$).

Assume that we can construct a unitary matrix $\tilde{Q}=[\tilde{\mathbf{q}}_1\,\tilde{\mathbf{q}}_2\,\cdots\,\tilde{\mathbf{q}}_n]$ such that $(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}=\min_i(\tilde{\mathbf{q}}_i^T B^{-1}\tilde{\mathbf{q}}_i)^{-1}<(\mathbf{u}^T B^{-1}\mathbf{u})^{-1}<\max_i(\tilde{\mathbf{q}}_i^T B^{-1}\tilde{\mathbf{q}}_i)^{-1}=(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}$. (For example, if $\max^-\neq\min^-$, then obtain $\tilde{Q}$ applying GS to $\mathbf{h}_{\max^-}$, $\mathbf{h}_i\ i\neq\max^-,\min^-$, and adding $\mathbf{r}_{\min^-}$ or applying GS to $\mathbf{r}_{\min^-}$, $\mathbf{r}_i\ i\neq\min^-,\max^-$, and adding $\mathbf{h}_{\max^-}$.) Then the idea is to consider $(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}$, $\tilde{\mathbf{q}}_\llcorner$, $(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}$, $\tilde{\mathbf{q}}_\lrcorner$, as approximations of λ_m, $\mathbf{q}_m$, λ_M, $\mathbf{q}_M$, respectively, and define $\tilde{C}_o$ so that

$$\tilde{C}_o\mathbf{u}=\sqrt{\frac{\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}}-\mathbf{u}^T B^{-1}\mathbf{u}}{\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}}-\frac{1}{(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}}}}\tilde{\mathbf{q}}_\llcorner+\sqrt{\frac{\mathbf{u}^T B^{-1}\mathbf{u}-\frac{1}{(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}}}{\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}}-\frac{1}{(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}}}}\tilde{\mathbf{q}}_\lrcorner.$$

Observe that $\|\tilde{Q}^T\tilde{C}_o\mathbf{u}\|^2=\mathbf{u}^T\tilde{C}_o^T\tilde{Q}\tilde{Q}^T\tilde{C}_o\mathbf{u}=1$, $\mathbf{u}^T\tilde{C}_o^T\tilde{\mathcal{Q}}_{B^{-1}}\tilde{C}_o\mathbf{u}=\mathbf{u}^T B^{-1}\mathbf{u}$, and

$$\mathbf{u}^T\tilde{C}_o^T\tilde{\mathcal{Q}}_{B^{-1}}^2\tilde{C}_o\mathbf{u}=(\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}}+\frac{1}{(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}})\mathbf{u}^T B^{-1}\mathbf{u}-\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}}$$

$$\ge\mathbf{u}^T\tilde{C}^T\tilde{\mathcal{Q}}_{B^{-1}}^2\tilde{C}\mathbf{u}=-\sum_{r\neq\llcorner,\lrcorner}[\tilde{Q}^T\tilde{C}\mathbf{u}]_r^2(\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}}-\frac{1}{(\tilde{\mathbf{q}}_r^T B^{-1}\tilde{\mathbf{q}}_r)^{-1}})(\frac{1}{(\tilde{\mathbf{q}}_r^T B^{-1}\tilde{\mathbf{q}}_r)^{-1}}-\frac{1}{(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}})$$

$$(\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}}+\frac{1}{(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}})\mathbf{u}^T B^{-1}\mathbf{u}-\frac{1}{(\tilde{\mathbf{q}}_\lrcorner^T B^{-1}\tilde{\mathbf{q}}_\lrcorner)^{-1}(\tilde{\mathbf{q}}_\llcorner^T B^{-1}\tilde{\mathbf{q}}_\llcorner)^{-1}}$$

for all $\tilde{C}$ such that $\mathbf{u}^T\tilde{C}^T\tilde{Q}\tilde{Q}^T\tilde{C}\mathbf{u}=1$, $\mathbf{u}^T\tilde{C}^T\tilde{\mathcal{Q}}_{B^{-1}}\tilde{C}\mathbf{u}=\mathbf{u}^T B^{-1}\mathbf{u}$.

Problem: $\mathbf{u}^T\tilde{C}_o^T QQ^T\tilde{C}_o\mathbf{u}=1$ ok, but $\mathbf{u}^T\tilde{C}_o^T B^{-1}\tilde{C}_o\mathbf{u}=\mathbf{u}^T B^{-1}\mathbf{u}$ not ok, $\mathbf{u}^T\tilde{C}_o^T B^{-2}\tilde{C}_o\mathbf{u}$ not maximum; so what inequalities the new actors satisfy in place of

$$\lambda_m\le\frac{\lambda_M\lambda_m}{\lambda_M+\lambda_m-(\mathbf{u}^T B^{-1}\mathbf{u})^{-1}}=\frac{1}{\frac{1}{\lambda_m}+\frac{1}{\lambda_M}-\frac{1}{\lambda_m\lambda_M\mathbf{u}^T B^{-1}\mathbf{u}}}=(\frac{\mathbf{u}^T C_o^T B^{-2}C_o\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1}=(\mathbf{r}_{C_o}^T B^{-1}\mathbf{r}_{C_o})^{-1}$$

$$\le(\frac{\mathbf{u}^T C^T B^{-2}C\mathbf{u}}{\mathbf{u}^T B^{-1}\mathbf{u}})^{-1}=(\mathbf{r}_C^T B^{-1}\mathbf{r}_C)^{-1}\le\mathbf{r}_C^T B\mathbf{r}_C=(\mathbf{u}^T B^{-1}\mathbf{u})^{-1}$$

for all C such that $\mathbf{u}^T C^T QQ^T C\mathbf{u}=1$, $\mathbf{u}^T C^T B^{-1}C\mathbf{u}=\mathbf{u}^T B^{-1}\mathbf{u}$?

Some details on Enlarge $\sigma(\mathcal{U}_B)$ and $\sigma((\mathcal{U}_{B^{-1}})^{-1})$. Let B be a real symmetric positive definite $n \times n$ matrix ($Q = [\mathbf{q}_1\, \mathbf{q}_2\, \dots\, \mathbf{q}_n]$, $Q^HQ = I$, $B\mathbf{q}_i = \lambda_i\mathbf{q}_i$, $\lambda_i > 0$, $i = 1 \dots n$). For $\mathbf{u} \in \mathbb{R}^n$, $\|\mathbf{u}\| = 1$, set

$$\mathbf{r} := \frac{B^{-1/2}\mathbf{u}}{\sqrt{\mathbf{u}^HB^{-1}\mathbf{u}}} \;\Rightarrow\; \min_i \lambda_i \le (\mathbf{r}^HB^{-1}\mathbf{r})^{-1} = (\frac{\mathbf{u}^HB^{-2}\mathbf{u}}{\mathbf{u}^HB^{-1}\mathbf{u}})^{-1} \le \mathbf{r}^HB\mathbf{r} = (\mathbf{u}^HB^{-1}\mathbf{u})^{-1},$$

$$\mathbf{h} := \frac{B^{1/2}\mathbf{u}}{\sqrt{\mathbf{u}^HB\mathbf{u}}} \;\Rightarrow\; (\mathbf{h}^HB^{-1}\mathbf{h})^{-1} = \mathbf{u}^HB\mathbf{u} \le \mathbf{h}^HB\mathbf{h} = \frac{\mathbf{u}^HB^2\mathbf{u}}{\mathbf{u}^HB\mathbf{u}} \le \max_i \lambda_i$$

(recall that $\mathbf{z}^HB\mathbf{z}\,\mathbf{z}^HB^{-1}\mathbf{z} \ge 1$ if $\|\mathbf{z}\| = 1$, with ?equality verified iff $\mathbf{z}$ is eigenvector of B?), thus $(\mathbf{u}^HB^{-1}\mathbf{u})^{-1} \in (\,(\mathbf{r}^HB^{-1}\mathbf{r})^{-1}\,,\,(\mathbf{h}^HB^{-1}\mathbf{h})^{-1}\,)$ and $\mathbf{u}^HB\mathbf{u} \in (\,\mathbf{r}^HB\mathbf{r}\,,\,\mathbf{h}^HB\mathbf{h}\,)$ whenever $\mathbf{u}$ is not eigenvector of B. Note that $\mathbf{u}^HB^{-1}\mathbf{u} \to \frac{\mathbf{u}^HB^{-2}\mathbf{u}}{\mathbf{u}^HB^{-1}\mathbf{u}}$ is one step of power method applied to B^{-1}, convergent to $1/\min_i \lambda_i$, and $\mathbf{u}^HB\mathbf{u} \to \frac{\mathbf{u}^HB^2\mathbf{u}}{\mathbf{u}^HB\mathbf{u}}$ is one step of power method applied to B, convergent to $\max_i \lambda_i$.

$U = [\mathbf{u}_1\, \mathbf{u}_2\, \dots\, \mathbf{u}_n]$, $U^HU = I$, $\mathcal{U} = \mathrm{sd}\,U = \{UDU^H :\ D \text{ diagonal matrices}\}$,
$\|B - \mathcal{U}_B\|_F = \min_{X \in \mathrm{sd}\,U} \|B - X\|_F$, $\|B^{-1} - \mathcal{U}_{B^{-1}}\|_F = \min_{X \in \mathrm{sd}\,U} \|B^{-1} - X\|_F$
$\mathcal{U}_B = U\,\mathrm{diag}\,(\mathbf{u}_i^HB\mathbf{u}_i)U^H$, $\sigma(\mathcal{U}_B) = \{\mathbf{u}_i^HB\mathbf{u}_i = (\mathbf{h}_i^HB^{-1}\mathbf{h}_i)^{-1}\} \subset \overline{\sigma(B)}$
$(\mathcal{U}_{B^{-1}})^{-1} = U\,\mathrm{diag}\,((\mathbf{u}_i^HB^{-1}\mathbf{u}_i)^{-1})U^H$, $\sigma((\mathcal{U}_{B^{-1}})^{-1}) = \{(\mathbf{u}_i^HB^{-1}\mathbf{u}_i)^{-1} = \mathbf{r}_i^HB\mathbf{r}_i\} \subset \overline{\sigma(B)}$,
Note: since $(\mathbf{u}_i^HB^{-1}\mathbf{u}_i)^{-1} \le \mathbf{u}_i^HB\mathbf{u}_i$, $\sigma((\mathcal{U}_{B^{-1}})^{-1})$ is shifted on the left with respect to $\sigma(\mathcal{U}_B)$

Definition of the indeces min and max (question: when min = max ?):
$\underline{(\mathbf{r}_{\min^-}^HB^{-1}\mathbf{r}_{\min^-})^{-1} \le} \mathbf{r}_{\min}^HB\mathbf{r}_{\min} \le \mathbf{r}_i^HB\mathbf{r}_i = (\mathbf{u}_i^HB^{-1}\mathbf{u}_i)^{-1} \le \mathbf{u}_i^HB\mathbf{u}_i = (\mathbf{h}_i^HB^{-1}\mathbf{h}_i)^{-1} \le \mathbf{h}_i^HB\mathbf{h}_i \le \mathbf{h}_{\max}^HB\mathbf{h}_{\max}$

(for the definition of $\min^-$, see below). Apply Gram-Schmidt to $\{\mathbf{r}_i\}$ starting from $\mathbf{r}_{\min}$:
$\mathbf{v}_{\min} = \mathbf{r}_{\min}$, $\underline{\mathbf{v}_{\min}^HB\mathbf{v}_{\min} = \mathbf{r}_{\min}^HB\mathbf{r}_{\min}}$, $t \ne \min$:
$\mathbf{v}_t = (\mathbf{r}_t - \mathbf{r}_t^H\mathbf{v}_{\min}\mathbf{v}_{\min})/\|\mathbf{r}_t - \mathbf{r}_t^H\mathbf{v}_{\min}\mathbf{v}_{\min}\|$, $\underline{\mathbf{v}_t^HB\mathbf{v}_t \ge \mathbf{r}_t^HB\mathbf{r}_t}$, , $s \ne \min, t$:
$\mathbf{v}_s = (\mathbf{r}_s - \mathbf{r}_s^H\mathbf{v}_{\min}\mathbf{v}_{\min} - \mathbf{r}_s^H\mathbf{v}_t\mathbf{v}_t)/\|\mathbf{r}_s - \mathbf{r}_s^H\mathbf{v}_{\min}\mathbf{v}_{\min} - \mathbf{r}_s^H\mathbf{v}_t\mathbf{v}_t\|$, $\underline{\mathbf{v}_s^HB\mathbf{v}_s \ge \mathbf{r}_s^HB\mathbf{r}_s}$
$\dots$ I have proved the above two inequalities (see the next page). Conjecture: the inequality is true also for the successive steps ....
$V = [\mathbf{v}_1\, \mathbf{v}_2\, \dots\, \mathbf{v}_n] \Rightarrow \mathbf{r}_{\min}^HB\mathbf{r}_{\min} = \min_i \lambda_i(\mathcal{V}_B) \le \lambda_i(\mathcal{U}_B)$ (?? $\lambda_i(\mathcal{U}_B) \le \max_i \lambda_i(\mathcal{V}_B)$??).
If max $\ne$ min: in G.S. choose $t, s, .. \ne \max$ and set $\tilde{V} = [\mathbf{v}_1 .. \mathbf{h}_{\max} .. \mathbf{v}_n]$ ($\tilde{V}\mathbf{e}_{\max} = \mathbf{h}_{\max}$) $\Rightarrow$ $\mathbf{r}_{\min}^HB\mathbf{r}_{\min} = \min_i \lambda_i(\tilde{\mathcal{V}}_B) \le \lambda_i(\mathcal{U}_B) \le \mathbf{h}_{\max}^HB\mathbf{h}_{\max} \le \max_i \lambda_i(\tilde{\mathcal{V}}_B)$.

Definition of the indeces $\min^-$ and $\max^-$ (question: when $\min^- = \max^-$?):
$(\mathbf{r}_{\min^-}^HB^{-1}\mathbf{r}_{\min^-})^{-1} \le (\mathbf{r}_i^HB^{-1}\mathbf{r}_i)^{-1} \le \mathbf{r}_i^HB\mathbf{r}_i = (\mathbf{u}_i^HB^{-1}\mathbf{u}_i)^{-1} \le \mathbf{u}_i^HB\mathbf{u}_i = (\mathbf{h}_i^HB^{-1}\mathbf{h}_i)^{-1} \le (\mathbf{h}_{\max^-}^HB^{-1}\mathbf{h}_{\max^-})^{-1} \underline{\le \mathbf{h}_{\max}^HB\mathbf{h}_{\max}}$

Apply Gram-Schmidt to $\{\mathbf{h}_i\}$ starting from $\mathbf{h}_{\max^-}$ (set $t := t^-$, $s := s^-$):
$\mathbf{w}_{\max^-} = \mathbf{h}_{\max^-}$, $\underline{\mathbf{w}_{\max^-}^HB^{-1}\mathbf{w}_{\max^-} = \mathbf{h}_{\max^-}^HB^{-1}\mathbf{h}_{\max^-}}$, $t \ne \max^-$:
$\mathbf{w}_t = (\mathbf{h}_t - \mathbf{h}_t^H\mathbf{w}_{\max^-}\mathbf{w}_{\max^-})/\|\mathbf{h}_t - \mathbf{h}_t^H\mathbf{w}_{\max^-}\mathbf{w}_{\max^-}\|$, $\underline{\mathbf{w}_t^HB^{-1}\mathbf{w}_t \ge \mathbf{h}_t^HB^{-1}\mathbf{h}_t}$, , $s \ne \max^-, t$:
$\mathbf{w}_s = (\mathbf{h}_s - \mathbf{h}_s^H\mathbf{w}_{\max^-}\mathbf{w}_{\max^-} - \mathbf{h}_s^H\mathbf{w}_t\mathbf{w}_t)/\|\mathbf{h}_s - \mathbf{h}_s^H\mathbf{w}_{\max^-}\mathbf{w}_{\max^-} - \mathbf{h}_s^H\mathbf{w}_t\mathbf{w}_t\|$, $\underline{\mathbf{w}_s^HB^{-1}\mathbf{w}_s \ge \mathbf{h}_s^HB^{-1}\mathbf{h}_s}$
$\dots$ I have proved the above two inequalities. Conjecture: the inequality is true also for the successive steps ....
$W = [\mathbf{w}_1\, \mathbf{w}_2\, \dots\, \mathbf{w}_n] \Rightarrow \lambda_i((\mathcal{U}_{B^{-1}})^{-1}) \le \max_i \lambda_i((\mathcal{W}_{B^{-1}})^{-1}) = (\mathbf{h}_{\max^-}^HB^{-1}\mathbf{h}_{\max^-})^{-1}$ (?$\min_i \lambda_i((\mathcal{W}_{B^{-1}})^{-1}) \le \lambda_i((\mathcal{U}_{B^{-1}})^{-1})$?).
If $\min^- \ne \max^-$: in G.S. choose $t, s, .. \ne \min^-$ and set $\tilde{W} = [\mathbf{w}_1 .. \mathbf{r}_{\min^-} .. \mathbf{w}_n]$ ($\tilde{W}\mathbf{e}_{\min^-} = \mathbf{r}_{\min^-}$) $\Rightarrow$
$\min_i \lambda_i((\tilde{\mathcal{W}}_{B^{-1}})^{-1}) \le (\mathbf{r}_{\min^-}^HB^{-1}\mathbf{r}_{\min^-})^{-1} \le \lambda_i((\mathcal{U}_{B^{-1}})^{-1}) \le \max_i \lambda_i((\tilde{\mathcal{W}}_{B^{-1}})^{-1}) = (\mathbf{h}_{\max^-}^HB^{-1}\mathbf{h}_{\max^-})^{-1}$.

On the two questions of the previous page

(1) Let us prove the two inequalities satisfied by the vectors $\mathbf{v}_t$, $\mathbf{v}_s$ generated by G.S.:

It can be shown that

$$\mathbf{v}_t^H B \mathbf{v}_t = \frac{\mathbf{r}_t^H B \mathbf{r}_t + (\mathbf{r}_t^H \mathbf{r}_{\min})^2 \mathbf{r}_{\min}^H B \mathbf{r}_{\min}}{1 - (\mathbf{r}_t^H \mathbf{r}_{\min})^2},$$

$$\mathbf{v}_s^H B \mathbf{v}_s = \frac{\mathbf{r}_s^H B \mathbf{r}_s + (\frac{\mathbf{r}_s^H \mathbf{r}_t - \mathbf{r}_t^H \mathbf{r}_{\min} \mathbf{r}_s^H \mathbf{r}_{\min}}{1-(\mathbf{r}_t^H \mathbf{r}_{\min})^2})^2 \mathbf{r}_t^H B \mathbf{r}_t + (\mathbf{r}_s^H \mathbf{r}_{\min} - \frac{\mathbf{r}_t^H \mathbf{r}_{\min} (\mathbf{r}_s^H \mathbf{r}_t - \mathbf{r}_t^H \mathbf{r}_{\min} \mathbf{r}_s^H \mathbf{r}_{\min})}{1-(\mathbf{r}_t^H \mathbf{r}_{\min})^2})^2 \mathbf{r}_{\min}^H B \mathbf{r}_{\min}}{1 - (\mathbf{r}_s^H \mathbf{r}_{\min})^2 - \frac{(\mathbf{r}_s^H \mathbf{r}_t - \mathbf{r}_t^H \mathbf{r}_{\min} \mathbf{r}_s^H \mathbf{r}_{\min})^2}{1-(\mathbf{r}_t^H \mathbf{r}_{\min})^2}}.$$

From the above explicit expressions of $\mathbf{v}_t^H B \mathbf{v}_t$ and $\mathbf{v}_s^H B \mathbf{v}_s$, it is obvious that $\mathbf{v}_t^H B \mathbf{v}_t \geq \mathbf{r}_t^H B \mathbf{r}_t$ and $\mathbf{v}_s^H B \mathbf{v}_s \geq \mathbf{r}_s^H B \mathbf{r}_s$. One can guess that also at any successive step of the G.S. procedure we have that

$$\mathbf{v}_k^H B \mathbf{v}_k = \frac{\mathbf{r}_k^H B \mathbf{r}_k + (\geq 0) \mathbf{r}_j^H B \mathbf{r}_j + \ldots + (\geq 0) \mathbf{r}_s^H B \mathbf{r}_s + (\geq 0) \mathbf{r}_t^H B \mathbf{r}_t + (\geq 0) \mathbf{r}_{\min}^H B \mathbf{r}_{\min}}{1 - (\geq 0) - (\geq 0) \ldots - (\geq 0) - (\geq 0)} \qquad (7)$$

which would imply the inequality $\mathbf{v}_k^H B \mathbf{v}_k \geq \mathbf{r}_k^H B \mathbf{r}_k$. But how to prove (7) ?

(2) If max $\neq$ min, then from U it can be constructed a unitary matrix $\tilde{V}$ such that

$$\min_i \lambda_i(\tilde{\mathcal{V}}_B) < \lambda_k(\mathcal{U}_B) < \max_i \lambda_i(\tilde{\mathcal{V}}_B), \quad \forall\ k$$

(see the previous page).
Question: when max = min ? Conjecture: if max = min, i.e. if $\exists\, \hat{s} \in \{1, \ldots, n\}$ |

$$\frac{\mathbf{u}_{\hat{s}}^H B^2 \mathbf{u}_{\hat{s}}}{\mathbf{u}_{\hat{s}}^H B \mathbf{u}_{\hat{s}}} \geq \frac{\mathbf{u}_s^H B^2 \mathbf{u}_s}{\mathbf{u}_s^H B \mathbf{u}_s}, \quad \mathbf{u}_{\hat{s}}^H B^{-1} \mathbf{u}_{\hat{s}} \geq \mathbf{u}_s^H B^{-1} \mathbf{u}_s, \quad s = 1, \ldots, n, \qquad (\#)$$

then the above inequalities must hold for any other $\hat{s} \in \{1, \ldots, n\}$; in other words the above inequalities can be satisfied only if they are all equalities:

$$\frac{\mathbf{u}_{\hat{s}}^H B^2 \mathbf{u}_{\hat{s}}}{\mathbf{u}_{\hat{s}}^H B \mathbf{u}_{\hat{s}}} = \frac{\mathbf{u}_s^H B^2 \mathbf{u}_s}{\mathbf{u}_s^H B \mathbf{u}_s}, \quad \mathbf{u}_{\hat{s}}^H B^{-1} \mathbf{u}_{\hat{s}} = \mathbf{u}_s^H B^{-1} \mathbf{u}_s, \quad s = 1, \ldots, n. \qquad (@)$$

If the Conjecture is true, then Problem: find all $U = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_n]$ unitary for which (@) hold.
See the next page

B Hermitian positive definite (Hpd) $n \times n$ matrix, $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n \in \mathbb{C}^n$ such that $\mathbf{u}_i^H \mathbf{u}_j = \delta_{ij}$.

Conjecture: if $\exists\, \hat{s} \in \{1, \ldots, n\}$ such that

$$\frac{\mathbf{u}_{\hat{s}}^H B^2 \mathbf{u}_{\hat{s}}}{\mathbf{u}_{\hat{s}}^H B \mathbf{u}_{\hat{s}}} \geq \frac{\mathbf{u}_s^H B^2 \mathbf{u}_s}{\mathbf{u}_s^H B \mathbf{u}_s}, \quad \mathbf{u}_{\hat{s}}^H B^{-1} \mathbf{u}_{\hat{s}} \geq \mathbf{u}_s^H B^{-1} \mathbf{u}_s, \quad s = 1, \ldots, n, \qquad (\#)$$

then the above inequalities must hold for any other $\hat{s} \in \{1, \ldots, n\}$; in other words the above inequalities can be satisfied only if they are all equalities:

$$\frac{\mathbf{u}_{\hat{s}}^H B^2 \mathbf{u}_{\hat{s}}}{\mathbf{u}_{\hat{s}}^H B \mathbf{u}_{\hat{s}}} = \frac{\mathbf{u}_s^H B^2 \mathbf{u}_s}{\mathbf{u}_s^H B \mathbf{u}_s}, \quad \mathbf{u}_{\hat{s}}^H B^{-1} \mathbf{u}_{\hat{s}} = \mathbf{u}_s^H B^{-1} \mathbf{u}_s, \quad s = 1, \ldots, n. \qquad (@)$$

Problem: If the Conjecture is true, then find all $U = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_n]$ unitary for which (@) hold.

The conjecture is true for $n = 2$: $B \in \mathbb{C}^{2\times 2}$ Hpd, $\mathbf{u}, \mathbf{v} \in \mathbb{C}^2$, $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$, $\mathbf{u}^H \mathbf{v} = 0 \Rightarrow \mathbf{u} = \alpha_1 \mathbf{q}_1 + \alpha_2 \mathbf{q}_2$, $\mathbf{v} = \beta_1 \mathbf{q}_1 + \beta_2 \mathbf{q}_2$ ($B\mathbf{q}_j = \lambda_j \mathbf{q}_j$, $\lambda_j > 0$, $\mathbf{q}_j^H \mathbf{q}_i = \delta_{ij}$, $i, j = 1, 2$), $|\alpha_1| + |\alpha_2| = |\beta_1| + |\beta_2| = 1$, $\overline{\alpha_1}\beta_1 + \overline{\alpha_2}\beta_2 = 0$. Thus we have

$$\frac{\mathbf{u}^H B^2 \mathbf{u}}{\mathbf{u}^H B \mathbf{u}} \geq \frac{\mathbf{v}^H B^2 \mathbf{v}}{\mathbf{v}^H B \mathbf{v}}, \quad \mathbf{u}^H B^{-1} \mathbf{u} \geq \mathbf{v}^H B^{-1} \mathbf{v}, \qquad (\#)$$

$$\Leftrightarrow \quad \frac{|\alpha_1|^2\lambda_1^2 + |\alpha_2|^2\lambda_2^2}{|\alpha_1|^2\lambda_1 + |\alpha_2|^2\lambda_2} \geq \frac{|\beta_1|^2\lambda_1^2 + |\beta_2|^2\lambda_2^2}{|\beta_1|^2\lambda_1 + |\beta_2|^2\lambda_2}, \quad |\alpha_1|^2\lambda_1^{-1} + |\alpha_2|^2\lambda_2^{-1} \geq |\beta_1|^2\lambda_1^{-1} + |\beta_2|^2\lambda_2^{-1} \quad \Leftrightarrow$$

$$\frac{\lambda_1^2 - |\alpha_2|^2(\lambda_1^2 - \lambda_2^2)}{\lambda_1 - |\alpha_2|^2(\lambda_1 - \lambda_2)} \geq \frac{\lambda_1^2 - |\beta_2|^2(\lambda_1^2 - \lambda_2^2)}{\lambda_1 - |\beta_2|^2(\lambda_1 - \lambda_2)}, \quad \lambda_1^{-1} - |\alpha_2|^2(\lambda_1^{-1} - \lambda_2^{-1}) \geq \lambda_1^{-1} - |\beta_2|^2(\lambda_1^{-1} - \lambda_2^{-1}) \quad \Leftrightarrow$$

$$(\lambda_1^2 - |\alpha_2|^2(\lambda_1^2 - \lambda_2^2))(\lambda_1 - |\beta_2|^2(\lambda_1 - \lambda_2)) \geq (\lambda_1^2 - |\beta_2|^2(\lambda_1^2 - \lambda_2^2))(\lambda_1 - |\alpha_2|^2(\lambda_1 - \lambda_2)),$$

$$|\alpha_2|^2(\lambda_2 - \lambda_1)/(\lambda_2\lambda_1) \leq |\beta_2|^2(\lambda_2 - \lambda_1)/(\lambda_2\lambda_1) \quad \Leftrightarrow$$

$$\lambda_1^3 - \lambda_1^2|\beta_2|^2(\lambda_1 - \lambda_2) - |\alpha_2|^2\lambda_1(\lambda_1^2 - \lambda_2^2) + |\alpha_2|^2|\beta_2|^2(\lambda_1 - \lambda_2)(\lambda_1^2 - \lambda_2^2) \geq$$
$$\lambda_1^3 - \lambda_1^2|\alpha_2|^2(\lambda_1 - \lambda_2) - |\beta_2|^2\lambda_1(\lambda_1^2 - \lambda_2^2) + |\beta_2|^2|\alpha_2|^2(\lambda_1 - \lambda_2)(\lambda_1^2 - \lambda_2^2), \ (|\alpha_2|^2 - |\beta_2|^2)(\lambda_2 - \lambda_1)/(\lambda_2\lambda_1) \leq 0 \quad \Leftrightarrow$$

$$-\lambda_1^2|\beta_2|^2(\lambda_1 - \lambda_2) - |\alpha_2|^2\lambda_1(\lambda_1^2 - \lambda_2^2) \geq -\lambda_1^2|\alpha_2|^2(\lambda_1 - \lambda_2) - |\beta_2|^2\lambda_1(\lambda_1^2 - \lambda_2^2), \quad \ldots \leq 0 \quad \Leftrightarrow$$

$$-\lambda_1^2(|\beta_2|^2 - |\alpha_2|^2)(\lambda_1 - \lambda_2) - (|\alpha_2|^2 - |\beta_2|^2)\lambda_1(\lambda_1^2 - \lambda_2^2) \geq 0, \quad \ldots \leq 0 \quad \Leftrightarrow$$

$$(|\beta_2|^2 - |\alpha_2|^2)\lambda_1\lambda_2(\lambda_1 - \lambda_2) \geq 0, \quad (|\beta_2|^2 - |\alpha_2|^2)(\lambda_1 - \lambda_2)/(\lambda_2\lambda_1) \leq 0$$

and the above inequalities hold only if they are equalities.

The case $n = 3$: $B \in \mathbb{C}^{3\times 3}$ Hpd, $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{C}^3$, $\|\mathbf{u}\| = \|\mathbf{v}\| = \|\mathbf{w}\| = 1$, $\mathbf{u}^H \mathbf{v} = \mathbf{u}^H \mathbf{w} = \mathbf{w}^H \mathbf{v} = 0 \Rightarrow \mathbf{u} = \alpha_1 \mathbf{q}_1 + \alpha_2 \mathbf{q}_2 + \alpha_3 \mathbf{q}_3$, $\mathbf{v} = \beta_1 \mathbf{q}_1 + \beta_2 \mathbf{q}_2 + \beta_3 \mathbf{q}_3$, $\mathbf{w} = \gamma_1 \mathbf{q}_1 + \gamma_2 \mathbf{q}_2 + \gamma_3 \mathbf{q}_3$ ($B\mathbf{q}_j = \lambda_j \mathbf{q}_j$, $\lambda_j > 0$, $\mathbf{q}_j^H \mathbf{q}_i = \delta_{ij}$, $i, j = 1, 2, 3$), $|\alpha_1|^2 + |\alpha_2|^2 + |\alpha_3|^2 = |\beta_1|^2 + |.|^2 + |.|^2 = |\gamma_1|^2 + |.|^2 + |.|^2 = 1$, $\overline{\alpha_1}\beta_1 + \overline{\alpha_2}\beta_2 + \overline{\alpha_3}\beta_3 = \overline{\alpha_1}\gamma_1 + . + . = \overline{\gamma_1}\beta_1 + . + . = 0$. Thus the conditions

$$\frac{\mathbf{u}^H B^2 \mathbf{u}}{\mathbf{u}^H B \mathbf{u}} \geq \frac{\mathbf{v}^H B^2 \mathbf{v}}{\mathbf{v}^H B \mathbf{v}}, \quad \frac{\mathbf{u}^H B^2 \mathbf{u}}{\mathbf{u}^H B \mathbf{u}} \geq \frac{\mathbf{w}^H B^2 \mathbf{w}}{\mathbf{w}^H B \mathbf{w}}, \quad \mathbf{u}^H B^{-1} \mathbf{u} \geq \mathbf{v}^H B^{-1} \mathbf{v}, \quad \mathbf{u}^H B^{-1} \mathbf{u} \geq \mathbf{w}^H B^{-1} \mathbf{w}, \qquad (\#)$$

are equivalent to

$$(|\beta_2|^2 - |\alpha_2|^2)[\lambda_1\lambda_2(\lambda_1 - \lambda_2)] + (|\beta_3|^2 - |\alpha_3|^2)[\lambda_1\lambda_3(\lambda_1 - \lambda_3)] + (|\alpha_2|^2|\beta_3|^2 - |\alpha_3|^2|\beta_2|^2)[(\lambda_1 - \lambda_3)(\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)] \geq 0,$$
$$(|\gamma_2|^2 - |\alpha_2|^2)[\lambda_1\lambda_2(\lambda_1 - \lambda_2)] + (|\gamma_3|^2 - |\alpha_3|^2)[\lambda_1\lambda_3(\lambda_1 - \lambda_3)] + (|\alpha_2|^2|\gamma_3|^2 - |\alpha_3|^2|\gamma_2|^2)[(\lambda_1 - \lambda_3)(\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)] \geq 0,$$
$$(|\beta_2|^2 - |\alpha_2|^2)(\lambda_1 - \lambda_2)/(\lambda_2\lambda_1) + (|\beta_3|^2 - |\alpha_3|^2)(\lambda_1 - \lambda_3)/(\lambda_3\lambda_1) \leq 0,$$
$$(|\gamma_2|^2 - |\alpha_2|^2)(\lambda_1 - \lambda_2)/(\lambda_2\lambda_1) + (|\gamma_3|^2 - |\alpha_3|^2)(\lambda_1 - \lambda_3)/(\lambda_3\lambda_1) \leq 0$$

Do the above inequalities hold only if they are equalities ?

Enrico Bozzo: The conjecture (that, in (#), the inequalities are satisfied only if they are equalities) is false, even for $n = 3$. Choose

$$B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \mathbf{u}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{u}_3 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

$\mathbf{u}_1^H B^{-1} \mathbf{u}_1 = \frac{2}{3}$, $\mathbf{r}_1 = \frac{1}{\sqrt{\mathbf{u}_1^H B^{-1} \mathbf{u}_1}} B^{-\frac{1}{2}} \mathbf{u}_1 = \sqrt{\frac{3}{2}} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{6}} \end{bmatrix}$, $\mathbf{r}_1^H B \mathbf{r}_1 = \frac{1}{\mathbf{u}_1^H B^{-1} \mathbf{u}_1} = \frac{3}{2}$.

$\mathbf{u}_2^H B^{-1} \mathbf{u}_2 = \frac{5}{9}$, $\mathbf{r}_2 = \frac{1}{\sqrt{\mathbf{u}_2^H B^{-1} \mathbf{u}_2}} B^{-\frac{1}{2}} \mathbf{u}_2 = \sqrt{\frac{3}{10}} \begin{bmatrix} 1 \\ \sqrt{2} \\ \frac{1}{\sqrt{3}} \end{bmatrix}$, $\mathbf{r}_2^H B \mathbf{r}_2 = \frac{1}{\mathbf{u}_2^H B^{-1} \mathbf{u}_2} = \frac{9}{5}$.

$\mathbf{u}_3^H B^{-1} \mathbf{u}_3 = \frac{11}{18}$, $\mathbf{r}_3 = \frac{1}{\sqrt{\mathbf{u}_3^H B^{-1} \mathbf{u}_3}} B^{-\frac{1}{2}} \mathbf{u}_3 = \sqrt{\frac{6}{11}} \begin{bmatrix} 1 \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}$, $\mathbf{r}_3^H B \mathbf{r}_3 = \frac{1}{\mathbf{u}_3^H B^{-1} \mathbf{u}_3} = \frac{18}{11}$.

$\mathbf{u}_1^H B \mathbf{u}_1 = 2$, $\mathbf{h}_1 = \frac{1}{\sqrt{\mathbf{u}_1^H B \mathbf{u}_1}} B^{\frac{1}{2}} \mathbf{u}_1 = \frac{1}{2} \begin{bmatrix} 1 \\ 0 \\ -\sqrt{3} \end{bmatrix}$, $\mathbf{h}_1^H B \mathbf{h}_1 = \frac{\mathbf{u}_1^H B^2 \mathbf{u}_1}{\mathbf{u}_1^H B \mathbf{u}_1} = \frac{5}{2}$

$\mathbf{u}_2^H B \mathbf{u}_2 = 2$, $\mathbf{h}_2 = \frac{1}{\sqrt{\mathbf{u}_2^H B \mathbf{u}_2}} B^{\frac{1}{2}} \mathbf{u}_2 = \frac{1}{\sqrt{12}} \begin{bmatrix} 1 \\ 2\sqrt{2} \\ \sqrt{3} \end{bmatrix}$, $\mathbf{h}_2^H B \mathbf{h}_2 = \frac{\mathbf{u}_2^H B^2 \mathbf{u}_2}{\mathbf{u}_2^H B \mathbf{u}_2} = \frac{13}{6}$

$\mathbf{u}_3^H B \mathbf{u}_3 = 2$, $\mathbf{h}_3 = \frac{1}{\sqrt{\mathbf{u}_3^H B \mathbf{u}_3}} B^{\frac{1}{2}} \mathbf{u}_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -\sqrt{2} \\ \sqrt{3} \end{bmatrix}$, $\mathbf{h}_3^H B \mathbf{h}_3 = \frac{\mathbf{u}_3^H B^2 \mathbf{u}_3}{\mathbf{u}_3^H B \mathbf{u}_3} = \frac{7}{3}$

It is clear that $\mathbf{r}_1^H B \mathbf{r}_1 < \mathbf{r}_i^H B \mathbf{r}_i$, $\forall\, i \neq 1$, and that $\mathbf{h}_1^H B \mathbf{h}_1 > \mathbf{h}_i^H B \mathbf{h}_i$, $\forall\, i \neq 1$. Thus, this example shows that in order to have that min = max it is not necessary that (#) are all equalities, i.e. my conjecture is false.

Since min = max (= 1), applying Gram-Schmidt to $\mathbf{r}_{\min}$, $\mathbf{h}_{\max}$, $\mathbf{r}_t$, ..., the vector $\mathbf{h}_{\max}$ is changed because it is not orthogonal to $\mathbf{r}_{\min}$ ($\mathbf{h}_1^H \mathbf{r}_1 = \frac{\sqrt{3}}{2}$), i.e. we obtain

$$\mathbf{z}_1 = \mathbf{r}_1, \ \mathbf{z}_2 = \frac{\mathbf{h}_1 - (\mathbf{h}_1^H \mathbf{r}_1)\mathbf{r}_1}{\|\mathbf{h}_1 - (\mathbf{h}_1^H \mathbf{r}_1)\mathbf{r}_1\|} = \begin{bmatrix} -\frac{1}{2} \\ 0 \\ -\frac{\sqrt{3}}{2} \end{bmatrix}$$

we are afraid that $\mathbf{z}_2^H B \mathbf{z}_2 < \mathbf{h}_1^H B \mathbf{h}_1$, in this case it would not be guaranteed anymore that $\mathbf{z}_2^H B \mathbf{z}_2$ is an upper bound for the $\mathbf{u}_i^H B \mathbf{u}_i$, $i = 1, 2, 3$. But $\mathbf{z}_2^H B \mathbf{z}_2$ turns out to be equal to $\frac{5}{2} = \mathbf{h}_1^H B \mathbf{h}_1$. Is this lucky or not ?

In general, if max = min, GS to $\mathbf{r}_{\min}$, $\mathbf{h}_{\max}$ ($\mathbf{h}_{\max}$ and $\mathbf{r}_{\min}$ indep unless $\mathbf{u}$ eigenvector), ... $\Rightarrow$

$\mathbf{z}_1 = \mathbf{r}_{\min}$, $\mathbf{z}_2 = \frac{\mathbf{h}_{\max} - (\mathbf{h}_{\max}^H \mathbf{r}_{\min})\mathbf{r}_{\min}}{\|\mathbf{h}_{\max} - (\mathbf{h}_{\max}^H \mathbf{r}_{\min})\mathbf{r}_{\min}\|}$, ...

$\mathbf{z}_2^H B \mathbf{z}_2 = \frac{(\mathbf{h}_{\max} - (\mathbf{h}_{\max}^H \mathbf{r}_{\min})\mathbf{r}_{\min})^H B (\mathbf{h}_{\max} - (\mathbf{h}_{\max}^H \mathbf{r}_{\min})\mathbf{r}_{\min})}{(\mathbf{h}_{\max} - (\mathbf{h}_{\max}^H \mathbf{r}_{\min})\mathbf{r}_{\min})^H (\mathbf{h}_{\max} - (\mathbf{h}_{\max}^H \mathbf{r}_{\min})\mathbf{r}_{\min})} = \frac{\mathbf{h}_{\max}^H B \mathbf{h}_{\max} - 2(\mathbf{h}_{\max}^H \mathbf{r}_{\min}) \mathbf{r}_{\min}^H B \mathbf{h}_{\max} + (\mathbf{h}_{\max}^H \mathbf{r}_{\min})^2 \mathbf{r}_{\min}^H B \mathbf{r}_{\min}}{1 - (\mathbf{h}_{\max}^H \mathbf{r}_{\min})^2}$

$\geq \mathbf{h}_{\max}^H B \mathbf{h}_{\max}$? i.e. $(\mathbf{h}_{\max}^H \mathbf{r}_{\min})^2 (\mathbf{r}_{\min}^H B \mathbf{r}_{\min} + \mathbf{h}_{\max}^H B \mathbf{h}_{\max}) - 2(\mathbf{h}_{\max}^H \mathbf{r}_{\min}) \mathbf{r}_{\min}^H B \mathbf{h}_{\max} \geq 0$? $\quad \mathbf{h}_1^H B \mathbf{r}_1 = \sqrt{3}$.

Problems (1), (2), (3)

Given $B \in \mathbb{C}^{n\times n}$ Hermitian (or real symmetric) positive definite (i.e. $B = Q\,\mathrm{diag}\,(\lambda_i,\, i=1,\dots,n)Q^H$ with Q unitary and $\lambda_i > 0$), look for information about the unitary (or real unitary) matrices $U = [\mathbf{u}_1\, \mathbf{u}_2\, \cdots\, \mathbf{u}_n]$ ($\mathbf{u}_i^H\mathbf{u}_j = \delta_{ij}\ i,j = 1,\dots,n$) such that, respectively, max = min, max = min$^-$, max$^-$ = min, max$^-$ = min$^-$, or, equivalently, for which there exists an index s such that, respectively,

max = min

$$\frac{\mathbf{u}_s^H B^2\mathbf{u}_s}{\mathbf{u}_s^H B\mathbf{u}_s} \geq \frac{\mathbf{u}_i^H B^2\mathbf{u}_i}{\mathbf{u}_i^H B\mathbf{u}_i},\ \forall\, i, \quad \mathbf{u}_s^H B^{-1}\mathbf{u}_s \geq \mathbf{u}_i^H B^{-1}\mathbf{u}_i,\ \forall\, i,$$

max = min$^-$

$$\frac{\mathbf{u}_s^H B^2\mathbf{u}_s}{\mathbf{u}_s^H B\mathbf{u}_s} \geq \frac{\mathbf{u}_i^H B^2\mathbf{u}_i}{\mathbf{u}_i^H B\mathbf{u}_i},\ \forall\, i, \quad \frac{\mathbf{u}_s^H B^{-2}\mathbf{u}_s}{\mathbf{u}_s^H B^{-1}\mathbf{u}_s} \geq \frac{\mathbf{u}_i^H B^{-2}\mathbf{u}_i}{\mathbf{u}_i^H B^{-1}\mathbf{u}_i},\ \forall\, i,$$

max$^-$ = min

$$\mathbf{u}_s^H B\mathbf{u}_s \geq \mathbf{u}_i^H B\mathbf{u}_i,\ \forall\, i, \quad \mathbf{u}_s^H B^{-1}\mathbf{u}_s \geq \mathbf{u}_i^H B^{-1}\mathbf{u}_i,\ \forall\, i,$$

max$^-$ = min$^-$

$$\mathbf{u}_s^H B\mathbf{u}_s \geq \mathbf{u}_i^H B\mathbf{u}_i,\ \forall\, i, \quad \frac{\mathbf{u}_s^H B^{-2}\mathbf{u}_s}{\mathbf{u}_s^H B^{-1}\mathbf{u}_s} \geq \frac{\mathbf{u}_i^H B^{-2}\mathbf{u}_i}{\mathbf{u}_i^H B^{-1}\mathbf{u}_i},\ \forall\, i.$$

(I think that we can assume B diagonal.)

(1) Find $U = [\mathbf{u}_1\, \cdots\, \mathbf{u}_n]$ such that max = min = max$^-$, i.e. for which

$$\exists\, s: \quad \frac{\mathbf{u}_s^H B^2\mathbf{u}_s}{\mathbf{u}_s^H B\mathbf{u}_s} \geq \frac{\mathbf{u}_i^H B^2\mathbf{u}_i}{\mathbf{u}_i^H B\mathbf{u}_i},\ \forall\, i, \quad \mathbf{u}_s^H B^{-1}\mathbf{u}_s \geq \mathbf{u}_i^H B^{-1}\mathbf{u}_i,\ \forall\, i \quad \mathbf{u}_s^H B\mathbf{u}_s \geq \mathbf{u}_i^H B\mathbf{u}_i,\ \forall\, i.$$

(Note that for the remaining U we can define $\mathcal{Z}$ such that $\overline{\sigma(\mathcal{U}_B)} \subset \overline{\sigma(\mathcal{Z}_B)}$.)

(2) Find $U = [\mathbf{u}_1\, \cdots\, \mathbf{u}_n]$ such that min = max$^-$ = min$^-$, i.e. for which

$$\exists\, s: \quad \mathbf{u}_s^H B^{-1}\mathbf{u}_s \geq \mathbf{u}_i^H B^{-1}\mathbf{u}_i,\ \forall\, i, \quad \mathbf{u}_s^H B\mathbf{u}_s \geq \mathbf{u}_i^H B\mathbf{u}_i,\ \forall\, i, \quad \frac{\mathbf{u}_s^H B^{-2}\mathbf{u}_s}{\mathbf{u}_s^H B^{-1}\mathbf{u}_s} \geq \frac{\mathbf{u}_i^H B^{-2}\mathbf{u}_i}{\mathbf{u}_i^H B^{-1}\mathbf{u}_i},\ \forall\, i.$$

(Note that for the remaining U we can define $\mathcal{Z}$ such that $\overline{\sigma((\mathcal{U}_{B^{-1}})^{-1})} \subset \overline{\sigma((\mathcal{Z}_{B^{-1}})^{-1})}$.)

(3) Find $U = [\mathbf{u}_1\, \cdots\, \mathbf{u}_n]$ such that max$^-$ = min, i.e. for which

$$\exists\, s: \quad \mathbf{u}_s^H B\mathbf{u}_s \geq \mathbf{u}_i^H B\mathbf{u}_i,\ \forall\, i, \quad \mathbf{u}_s^H B^{-1}\mathbf{u}_s \geq \mathbf{u}_i^H B^{-1}\mathbf{u}_i,\ \forall\, i,$$

(Note that for the remaining U we can define $\mathcal{Z}$ such that $\overline{\sigma(\mathcal{U}_B)} \subset \overline{\sigma(\mathcal{Z}_B)}$ and $\overline{\sigma((\mathcal{U}_{B^{-1}})^{-1})} \subset \overline{\sigma((\mathcal{Z}_{B^{-1}})^{-1})}$ simultaneously.)

(I think that (1) and (2) are equivalent.)

Given $\hat{\mathbf{x}} \in \mathbb{R}^n$ well defined (for instance $\hat{\mathbf{x}}$ could be the solution of a linear system $A\mathbf{x} = \mathbf{b}$) such that the product $M\hat{\mathbf{x}}$ for some fixed M real symmetric positive definite is much more easily computable than computing $\hat{\mathbf{x}}$ (for instance, if $\hat{\mathbf{x}} = A^{-1}\mathbf{b}$, M could be chosen as A^TA or even as A, in case A is real symmetric positive definite), find a sequence $\mathbf{x}_k \in \mathbb{R}^n$ convergent to $\hat{\mathbf{x}}$, in the sense that $\|\hat{\mathbf{x}} - \mathbf{x}_k\|_M \to 0$, such that the computation of $\mathbf{x}_{k+1}$ from $\mathbf{x}_k$ is much more easy than computing $\hat{\mathbf{x}}$. (Recall that $\|\mathbf{v}\|_M^2 = (\mathbf{v}, \mathbf{v})_M$ where $(\mathbf{u}, \mathbf{v})_M = \mathbf{u}^T M \mathbf{v}$.)

Choose an arbitrary $\mathbf{x}_0 \in \mathbb{R}^n$. For $k = 0, 1, 2, \ldots$

compute the *residual* $M(\hat{\mathbf{x}} - \mathbf{x}_k)$; if it is the null vector, then $\mathbf{x}_k = \hat{\mathbf{x}}$, otherwise

$$\{ \text{ choose } s_k \text{ such that } (M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_k} \neq 0 \ ,$$

$$\text{set } \ \mathbf{x}_{k+1} = \mathbf{x}_k + \omega_k \mathbf{e}_{s_k}, \ \text{ where } \ \omega_k = \frac{(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_k}}{M_{s_k, s_k}} \ \}.$$

Note that ω_k satisfies the following three equivalent properties:

$$\|\hat{\mathbf{x}} - \mathbf{x}_k\|_M^2 - \frac{(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_k}^2}{M_{s_k, s_k}} = \|\hat{\mathbf{x}} - \mathbf{x}_k\|_M^2 - \|\omega_k \mathbf{e}_{s_k}\|_M^2 = \|\hat{\mathbf{x}} - \mathbf{x}_k - \omega_k \mathbf{e}_{s_k}\|_M^2 \leq \|\hat{\mathbf{x}} - \mathbf{x}_k - \omega \mathbf{e}_{s_k}\|_M^2, \ \forall \omega \in \mathbb{R} \, ;$$

$$(M(\hat{\mathbf{x}} - \mathbf{x}_{k+1}))_{s_k} = (\hat{\mathbf{x}} - \mathbf{x}_{k+1}, \mathbf{e}_{s_k})_M = (\hat{\mathbf{x}} - \mathbf{x}_k - \omega_k \mathbf{e}_{s_k}, \mathbf{e}_{s_k})_M = 0 \, ;$$

$$F(\mathbf{x}_k + \omega_k \mathbf{e}_{s_k}) \leq F(\mathbf{x}_k + \omega \mathbf{e}_{s_k}), \ \ \forall \omega \in \mathbb{R}, \ \text{ where } \ F(\mathbf{z}) = \frac{1}{2} \mathbf{z}^T M \mathbf{z} - \mathbf{z}^T M \hat{\mathbf{x}} \, ,$$

thus the s_k entry of $\mathbf{x}_k$ is changed so that the s_k entry of the new residual is zero.

Then $\lim_{k \to +\infty} \dfrac{|(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_k}|}{\sqrt{M_{s_k, s_k}}} = 0 \quad \Rightarrow$

$$\left[\ \lim_{k \to +\infty} |(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_k}| = 0. \ \right] \qquad \Rightarrow$$

Let $c \in (0, 1]$ be fixed. If, for each k, s_k is chosen such that $|(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_k}| \geq c |(M(\hat{\mathbf{x}} - \mathbf{x}_k))_i| \ \ \forall i$ (i.e. if, for each k, a quite big residual entry is made equal to zero), then $\mathbf{x}_k \ \to \ \hat{\mathbf{x}}$.

Moreover, $\forall \ k \, , \ \forall t : \ 1 \leq t \leq k$ we have

$$\hat{\mathbf{x}} - \mathbf{x}_k = \hat{\mathbf{x}} - \mathbf{x}_{k-t+1} - \sum_{j=1}^{t-1} \omega_{k-j} \mathbf{e}_{s_{k-j}} \ \Rightarrow \ \ (\hat{\mathbf{x}} - \mathbf{x}_k, \mathbf{e}_{s_{k-t}})_M = - \sum_{j=1}^{t-1} \omega_{k-j} (\mathbf{e}_{s_{k-j}}, \mathbf{e}_{s_{k-t}})_M$$

$$\Rightarrow \quad \frac{|(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_{k-t}}|}{\sqrt{M_{s_{k-t}, s_{k-t}}}} \leq \sum_{j=1}^{t-1} \frac{|(M(\hat{\mathbf{x}} - \mathbf{x}_{k-j}))_{s_{k-j}}|}{\sqrt{M_{s_{k-j}, s_{k-j}}}} \quad \Rightarrow$$

$$\left[\ |(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_{k-t}}| \leq \sqrt{\frac{\max_i M_{ii}}{\min_i M_{ii}}} \sum_{j=1}^{t-1} |(M(\hat{\mathbf{x}} - \mathbf{x}_{k-j}))_{s_{k-j}}|, \ \ \forall \ k \, , \ \forall t : \ 1 \leq t \leq k \, . \ \right] \quad \Rightarrow$$

Let $m \in \mathbb{N}$, $m \geq n$, be fixed. If, for each k and $r \in \{1, 2, \ldots, n\}$, exists $t = t_{k,r} \in \{1, \ldots, m\}$ such that $s_{k-t} = r$ (i.e. if all the residual entries are periodically made equal to zero), then $\mathbf{x}_k \ \to \ \hat{\mathbf{x}}$.

Proof: $\forall r$ we have $|(M(\hat{\mathbf{x}} - \mathbf{x}_k))_r| = |(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_{k-t}}| \leq \ldots \leq \ldots \sum_{j=1}^{m-1} \ldots \ \to \ 0$

Corollaries:

Case $\hat{\mathbf{x}} = A^{-1}\mathbf{b}$, $M = A$:=real symmetric positive definite ($M(\hat{\mathbf{x}} - \mathbf{x}_k) = \mathbf{b} - A\mathbf{x}_k$):
For $k = 0, 1, 2, \ldots$

$$\text{compute } \mathbf{b} - A\mathbf{x}_k \text{ ; if it is the null vector, then } \mathbf{x}_k = \hat{\mathbf{x}}, \text{ otherwise}$$

$$\{ \text{ choose } s_k \text{ such that } (\mathbf{b} - A\mathbf{x}_k)_{s_k} \neq 0 \text{ ,}$$

$$\text{set } \mathbf{x}_{k+1} = \mathbf{x}_k + \omega_k \mathbf{e}_{s_k}, \text{ where } \omega_k = \frac{(\mathbf{b} - A\mathbf{x}_k)_{s_k}}{A_{s_k,s_k}} \text{ }\}.$$

Let $c \in (0, 1]$ be fixed. If, for each k, s_k is chosen such that $|(\mathbf{b} - A\mathbf{x}_k)_{s_k}| \geq c|(\mathbf{b} - A\mathbf{x}_k)_i| \ \forall\, i$, then $\mathbf{x}_k \to \hat{\mathbf{x}}$.
Let $m \in \mathbb{N}$, $m \geq n$, be fixed. If, for each k and $r \in \{1, 2, \ldots, n\}$, exists $t = t_{k,r} \in \{1, \ldots, m\}$ such that $s_{k-t} = r$, then $\mathbf{x}_k \to \hat{\mathbf{x}}$. Note: for $m = n$ and $s_k = k \bmod n + 1$ the latter method is the Gauss-Seidel method or, more precisely, $\mathbf{x}_n$, $\mathbf{x}_{2n}$, ... turn out to coincide with the first, second, ..., iterates generated by the Gauss-Seidel method (started with $\mathbf{x}_0$). So we have obtained an alternative proof of the convergence of the Gauss-Seidel method when applied to real symmetric positive definite linear systems.

Case $\hat{\mathbf{x}} = A^{-1}\mathbf{b}$, $M = A^T A$, A real non singular ($M(\hat{\mathbf{x}} - \mathbf{x}_k) = A^T(\mathbf{b} - A\mathbf{x}_k)$):
For $k = 0, 1, 2, \ldots$

$$\text{compute } A^T(\mathbf{b} - A\mathbf{x}_k) \text{ ; if it is the null vector, then } \mathbf{x}_k = \hat{\mathbf{x}}, \text{ otherwise}$$

$$\{ \text{ choose } s_k \text{ such that } (A^T(\mathbf{b} - A\mathbf{x}_k))_{s_k} \neq 0 \text{ ,}$$

$$\text{set } \mathbf{x}_{k+1} = \mathbf{x}_k + \omega_k \mathbf{e}_{s_k}, \text{ where } \omega_k = \frac{(A^T(\mathbf{b} - A\mathbf{x}_k))_{s_k}}{(A^T A)_{s_k,s_k}} \text{ }\}.$$

Let $c \in (0, 1]$ be fixed. If, for each k, s_k is chosen such that $|(A^T(\mathbf{b} - A\mathbf{x}_k))_{s_k}| \geq c|(A^T(\mathbf{b} - A\mathbf{x}_k))_i|$ $\forall\, i$, then $\mathbf{x}_k \to \hat{\mathbf{x}}$.
Let $m \in \mathbb{N}$, $m \geq n$, be fixed. If, for each k and $r \in \{1, 2, \ldots, n\}$, exists $t = t_{k,r} \in \{1, \ldots, m\}$ such that $s_{k-t} = r$, then $\mathbf{x}_k \to \hat{\mathbf{x}}$.

Question 1
Can we assume $M_{ii} = 1 \ \forall\, i$? For instance, can every real symmetric positive definite linear system be reduced to a real symmetric positive definite linear system $A\mathbf{x} = \mathbf{b}$ where $A_{ii} = 1 \ \forall\, i$? or, can every real normal linear system be reduced to a real normal linear system $A^T A\mathbf{x} = A^T\mathbf{b}$ where $[A^T A]_{ii} = 1 \ \forall\, i$?

Question 2
In the original Southwell method the parameter ω_k is only an approximation of the ratio $(M(\hat{\mathbf{x}} - \mathbf{x}_k))_{s_k}/M_{s_k,s_k}$, chosen so simple to make the updating of $(\mathbf{x}_k)_{s_k}$ (and thus of $\mathbf{x}_k$) computable by a hand desk calculator. Thus the s_k entry of $\mathbf{x}_k$ is changed so that the s_k entry of the new residual is *almost* zero. Try to prove the convergence of such old method.

Minimizzazione globale alla Gross

Risolvere il sistema lineare $A\mathbf{x} = \mathbf{b}$ è equivalente a cercare i punti estremali della funzione

$$h(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} - 2\mathbf{b}^T\mathbf{x}. \qquad (quadr)$$

Infatti, il sistema delle derivate parziali di tale funzione, che determina appunto i suoi estremali, è proprio $h_x(\mathbf{x}) = 2A\mathbf{x} - 2\mathbf{b} = \mathbf{0}$. In particolare, se A è definita positiva, allora l'unica soluzione $\overline{\mathbf{x}}$ del sistema lineare $A\mathbf{x} = \mathbf{b}$, corrisponde all'unico punto estremale per h, che è un punto di minimo assoluto. Quindi, $\overline{\mathbf{x}}$ potrà essere calcolata applicando alla funzione (quadr) un algoritmo per la minimizzazione non vincolata di funzioni, anziché i noti metodi per la risoluzione di sistemi lineari (e.g. Cholesky, Gradiente coniugato).

Nel seguito studiamo due tecniche (tt) e (tp), la seconda delle quali di più facile implementazione, per il calcolo dei punti di minimo $\overline{\mathbf{x}}$ di una funzione h generica. Lo scopo di tali tecniche è produrre una successione $\{\mathbf{x}_k\}$ i cui punti di accumulazione (quando esistono) siano punti stazionari per h. Inoltre, $\{\mathbf{x}_k\}$ deve essere una successione minimizzante h, i.e.

$$h(\mathbf{x}_{k+1}) < h(\mathbf{x}_k), \quad \text{for all } k.$$

Possibilmente, ma non necessariamente, se applicate a forme quadratiche (quadr) definite positive, le tecniche (tt) e (tp) dovrebbero produrre $\overline{\mathbf{x}}$ con un numero finito di passi, cioè si dovrebbe avere (in aritmetica esatta) $\mathbf{x}_m = \overline{\mathbf{x}}$ per qualche $m > 0$. I metodi per la minimizzazione di funzioni che rientrano in queste due tecniche sono quindi metodi *iterativi* e possono diventare diretti solo se applicati a funzioni quadratiche.

Le due tecniche (tt) e (tp) sono presentate con i rispettivi risultati di convergenza (globale) che le riguardano. Dal teorema di convergenza per la tecnica (tt), applicato alla forma (quadr), seguono noti risultati di convergenza dei metodi iterativi (per equazioni) lineari. Da quello per la tecnica (tp) segue un risultato di convergenza globale del metodo di Newton per il calcolo degli zeri di un sistema non lineare $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, modificato in modo tale che la successione $\{\mathbf{f}(\mathbf{x}_k)^T\mathbf{f}(\mathbf{x}_k)\}$ da esso generata sia decrescente. Infine, si ottiene banalmente la convergenza del metodo di discesa più ripida (o del Gradiente), usando sia la tecnica (tt) che quella (tp).

Due richiami sul Calcolo: i) Se $h \in C^1$ in un insieme contenente il segmento che unisce il punto $\mathbf{x} = (x_1 \cdots x_n)$ al punto $\mathbf{y} = (y_1 \cdots y_n)$, allora su tale segmento esiste un punto $\xi = (\xi_1 \cdots \xi_n)$, $\xi_i = y_i + \theta(x_i - y_i)$, $\theta \in (0,1)$, tale che

$$\begin{aligned} &h(x_1, \ldots, x_n) - h(y_1, \ldots, y_n) = h(\mathbf{x}) - h(\mathbf{y}) \\ &= h_x(\xi)(\mathbf{x} - \mathbf{y}) = \textstyle\sum_{i=1}^n \frac{\partial h}{\partial x_i}(\xi_1, \ldots, \xi_n)(x_i - y_i). \end{aligned} \qquad (TeoMedia)$$

ii) Se $\mathbf{x}(t)$ è una curva parametrica di classe C^1 per $t \in \mathcal{I}$ ed $h \in C^1$ in un insieme contenente il tratto $\mathbf{x}(t)$, $t \in \mathcal{I}$, allora

$$\frac{d}{dt}h(\mathbf{x}(t)) = \sum_{i=1}^{n} \frac{\partial h}{\partial x_i}(\mathbf{x}(t))\frac{dx_i(t)}{dt} = h_{\mathbf{x}}(\mathbf{x}(t))\mathbf{x}(t), \ t \in \mathcal{I} \qquad (DsuCurva)$$

... e sulla notazione: Nel seguito, se $\mathbf{v}$ è un vettore, allora il simbolo $\|\mathbf{v}\|$ indicherà la norma euclidea di $\mathbf{v}$, cioè $\|\mathbf{v}\| = \sqrt{\sum v_i^2}$. Consideremo, invece, anche norme di matrici diverse dalla norma spettrale (che è, ricordiamo, quella indotta dalla norma vettoriale euclidea).

Supponiamo che una funzione $h(\mathbf{x})$ di n variabili abbia un punto di minimo locale $\overline{\mathbf{x}}$. Necessariamente si dovrà avere

$$\frac{\partial h}{\partial x_j}(\overline{\mathbf{x}}) = 0, \quad j = 1, \ldots, m,$$

o, in forma compatta,

$$h_x(\overline{\mathbf{x}}) = \mathbf{0}. \tag{$SistDP$}$$

Sia $\mathbf{x}_0 \in \mathbb{R}^n$ una approssimazione di un $\overline{\mathbf{x}}$. Una componente connessa della curva di livello $h(\mathbf{x}) = h(\mathbf{x}_0)$ contiene $\overline{\mathbf{x}}$. Introducendo un vettore $\mathbf{s}_0$ di norma 1, che formi un angolo acuto con il vettore gradiente $h_{\mathbf{x}}(\mathbf{x}_0)$, e scegliendo un numero $\lambda_0 > 0$ opportuno, si può trovare un punto $\mathbf{x}_0 - \lambda_0 \mathbf{s}_0$ dove la funzione h è più piccola che nel punto $\mathbf{x}_0$. È naturale quindi definire $\mathbf{x}_1$, la nuova approssimazione di $\overline{\mathbf{x}}$, come segue:

$$\mathbf{x}_1 = \mathbf{x}_0 - \lambda_0 \mathbf{s}_0.$$

In ogni tecnica di minimizzazione la nuova approssimazione $\mathbf{x}_{k+1}$, del punto di minimo $\overline{\mathbf{x}}$, è definita dalla vecchia, $\mathbf{x}_k$, con questo procedimento. La scelta della *direzione di ricerca* $\mathbf{s}_k$ e del *passo* λ_k distinguerà una tecnica dalle altre.

In entrambe le tecniche (tt) e (tp) da noi trattate, le direzioni di ricerca $\mathbf{s}_k$ si scelgono tali che

$$\frac{h_{\mathbf{x}}(\mathbf{x}_k)\mathbf{s}_k}{\|h_x(\mathbf{x}_k)\|} \geq \gamma > 0. \tag{dir}$$

Si richiede, cioè, che l'angolo con il vettore gradiente di h (direzione di massima crescita per h) sia acuto uniformemente nell'intero processo di minimizzazione. Le due tecniche differiscono nella definizione del passo λ_k: nella prima, per scendere, si sfrutta per intero la potenzialità della direzione $\mathbf{s}_k$ scelta; nella seconda si cerca di definire un λ_k che produca una sufficiente decrescita ad un costo computazionale minimo. Tuttavia, entrambi i λ_k proposti in (tt) ed in (tp) soddisfano la condizione

$$\frac{\lambda_k}{\|h_x(\mathbf{x}_k)\|} \geq \sigma > 0. \tag{$passo$}$$

Quindi, si permette che il passo possa diventare piccolo solo quando si è prossimi alla soluzione $\overline{\mathbf{x}}$ (ovvero solo quando $\|h_x(\mathbf{x}_k)\|$ è piccolo).

Le due ipotesi (dir) e (passo) assicurano che i punti di accumulazione della successione $\mathbf{x}_k$ siano punti stazionari per h. Per avere la convergenza dell'intera successione $\{\mathbf{x}_k\}$, è ovviamente anche necessario che il passo λ_k non possa essere troppo lungo in prossimità della soluzione.

Matematica:

Stabiliamo, innanzitutto, il seguente Lemma fondamentale.

Lemma 1. *Sia $\mathcal{K}$ un insieme di $\mathbb{R}^n$ compatto e convesso dove la funzione h è di classe C^1 e tale che*

$$\|h_x(\mathbf{x})\| \geq M > 0. \tag{$gradlowb$}$$

Siano χ, γ numeri reali fissati tali che $0 < \chi < 1$, $0 < \gamma \leq 1$. Allora esiste $\mu = \mu(M, \gamma, \chi) > 0$ tale che

$$h(\mathbf{x} - \lambda \mathbf{s}) < h(\mathbf{x}) - \lambda \chi h_x(\mathbf{x})\mathbf{s} \leq h(\mathbf{x}) - \lambda \chi \gamma \|h_x(\mathbf{x})\|$$

per tutti gli $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{s} \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$ *tali che*

$$\mathbf{x} \in \mathcal{K},\ \|\mathbf{s}\| = 1,\ \frac{h_x(\mathbf{x})\mathbf{s}}{\|h_x(\mathbf{x})\|} \geq \gamma, \quad 0 < \lambda \leq \mu \quad e \quad \mathbf{x} - \lambda\mathbf{s} \in \mathcal{K}.$$

Dim. Esiste ν, $0 < \nu < \lambda$, tale che

$$h(\mathbf{x} - \lambda\mathbf{s}) - h(\mathbf{x}) = -\lambda h_x(\mathbf{x} - \nu\mathbf{s})\mathbf{s}.$$

Esiste, inoltre, $\mu = \mu(M, \gamma, \chi) > 0$ tale che

$$\nu < \mu \quad \Rightarrow \quad \|h_x(\mathbf{x}) - h_x(\mathbf{x} - \nu\mathbf{s})\| < (1 - \chi)M\gamma.$$

Quindi, se $\lambda \leq \mu$, allora

$$\begin{aligned} &h(\mathbf{x} - \lambda\mathbf{s}) - h(\mathbf{x}) + \lambda\chi h_x(\mathbf{x})\mathbf{s} \\ &\quad = \lambda[(h_x(\mathbf{x}) - h_x(\mathbf{x} - \nu\mathbf{s}))\mathbf{s} - (1 - \chi)h_x(\mathbf{x})\mathbf{s}] \\ &\quad = \lambda[\|h_x(\mathbf{x}) - h_x(\mathbf{x} - \nu\mathbf{s})\| - (1 - \chi)\gamma M] < 0. \end{aligned}$$

Teorema 1 (tt). *Sia* h *continua con le sue derivate nell'insieme* $\mathcal{I}_0 = \{\mathbf{x} : h(\mathbf{x}) \leq h(\mathbf{x}_0)\}$ *e* $\mathcal{I}_0$ *compatto. Siano* $\mathbf{s}_k \in \mathbb{R}^n$, ρ_k, $\lambda_k \in \mathbb{R}$ *tali che*

$$\|\mathbf{s}_k\| = 1,\ \frac{h_x(\mathbf{x}_k)\mathbf{s}_k}{\|h_x(\mathbf{x}_k)\|} \geq \gamma > 0,\ \sigma_k = \frac{\rho_k}{\|h_x(\mathbf{x}_k)\|} \geq \sigma > 0,$$

$$h(\mathbf{x}_k - \lambda_k\mathbf{s}_k) \leq h(\mathbf{x}_k - \lambda\mathbf{s}_k), \quad \forall\lambda \in (0, \rho_k].$$

Si ponga $\mathbf{x}_{k+1} = \mathbf{x}_k - \lambda_k\mathbf{s}_k$. *Allora*

$$\mathbf{x}_k \in \mathcal{I}_0, \quad h(\mathbf{x}_{k+1}) < h(\mathbf{x}_k), \quad \lim_{k\to\infty} h_x(\mathbf{x}_k) = \mathbf{0}.$$

Quindi, ogni punto di accumulazione di $\mathbf{x}_k$ *(Nota: esistono!) è soluzione del sistema delle derivate parziali di* h.

Commento 1. Se h ha più di un punto stazionario in $\mathcal{I}_0$ allora la successione $\{\mathbf{x}_k\}$ generata con la tecnica (tt) potrebbe non convergere. In realtà, la non convergenza si potrebbe avere solo nel caso critico in cui la funzione h assume lo stesso valore in almeno due distinti suoi punti stazionari in $\mathcal{I}_0$ (questo perchè $\{h(\mathbf{x}_k)\}$ è convergente!); in tal caso, infatti, $\mathbf{s}_k$ e λ_k potrebbero essere definiti, rispettando le ipotesi, in modo che $\mathbf{x}_k$ salti da uno all'altro. Se nei punti stazionari in $\mathcal{I}_0$, h assume valori diversi, allora $\mathbf{x}_k$ necessariamente convergerà ad uno di questi.

Nella tecnica di minimizzazione (tt) illustrata nel Teorema 1, il calcolo del passo λ_k richiede l'applicazione di un metodo iterativo per la minimizzazione della funzione unidimensionale $h(\mathbf{x}_k - \lambda\mathbf{s}_k)$ nell'insieme $(0, \rho_k]$.

È possibile introdurre una tecnica (tp) più pratica per la definizione di λ_k. Questa è illustrata nel seguente Teorema 2.

Teorema 2 (tp). *Sia* h *continua con le sue derivate nell'insieme* $\mathcal{I}_0 = \{\mathbf{x} : h(\mathbf{x}) \leq h(\mathbf{x}_0)\}$ *e* $\mathcal{I}_0$ *compatto. Sia* $\{c_i\}$ *una successione di numeri reali tale che*

$$c_0 = 1,\ 0 < c_i < c_{i-1},\ \lim_{i\to\infty} c_i = 0,\ \frac{c_{i-1}}{c_i} \leq c. \qquad (succ)$$

Per ogni $k \geq 0$, siano $\mathbf{s}_k \in \mathbb{R}^n$, ρ_k, $\lambda_k \in \mathbb{R}$ tali che

$$\|\mathbf{s}_k\| = 1, \ \frac{h_x(\mathbf{x}_k)\mathbf{s}_k}{\|h_x(\mathbf{x}_k)\|} \geq \gamma > 0, \ \infty > \sigma_k = \frac{\rho_k}{\|h_x(\mathbf{x}_k)\|} \geq \sigma > 0,$$

$$\begin{array}{l} \lambda_k = c_j\rho_k \ \textit{dove} \\ h(\mathbf{x}_k - c_j\rho_k\mathbf{s}_k) < h(\mathbf{x}_k) - c_j\rho_k\chi h_x(\mathbf{x}_k)\mathbf{s}_k \qquad (AG1) \\ h(\mathbf{x}_k - c_i\rho_k\mathbf{s}_k) \geq h(\mathbf{x}_k) - c_i\rho_k\chi h_x(\mathbf{x}_k)\mathbf{s}_k, \ i = 0, \ldots, j-1. \end{array}$$

Si ponga $\mathbf{x}_{k+1} = \mathbf{x}_k - \lambda_k\mathbf{s}_k$. Allora

$$\mathbf{x}_k \in \mathcal{I}_0, \quad h(\mathbf{x}_{k+1}) < h(\mathbf{x}_k), \quad \lim_{k\to\infty} h_x(\mathbf{x}_k) = \mathbf{0}.$$

Quindi, ogni punto di accumulazione di $\mathbf{x}_k$ (Nota: esistono!) è soluzione del sistema delle derivate parziali di h.

Commento 2. Quanto si è osservato sul Teorema 1 (vedi Commento 1) vale anche per il Teorema 2. Inoltre:

Per entrambe le scelte di λ_k proposte nei Teoremi 1 e 2 si può dimostrare che l'ipotesi su ρ_k implica la condizione $\frac{\lambda_k}{\|h_x(\mathbf{x}_k)\|} \geq \sigma' > 0$ (se $h \in C^2$), necessaria per la convergenza.

(Problema: l'ipotesi su ρ_k implica la seconda condizione di AG o, per lo meno, la disuguaglianza $(h_x(\mathbf{x}_{k+1}) - h_x(\mathbf{x}_k))(\mathbf{x}_{k+1} - \mathbf{x}_k) > 0$?)

La definizione di λ_k suggerita nel Teorema 2 (tp) utilizza una procedura, di tipo "backtracking", sicuramente di più facile implementazione di quella suggerita nel Teorema 1 (tt). Ad ogni iterazione k, per $i = 0, \ldots$ si pone $\lambda_k = c_i\rho_k$, con c_i definita in (succ), finché la disuguaglianza (AG1) è verificata (poiché inizialmente si prova sempre la scelta $\lambda_k = \rho_k$, il numero ρ_k non potrà essere uguale a $+\infty$ per nessun k). Si noti che la condizione $c_{i-1}/c_i \leq c$ ($\Rightarrow c_i \to 0$ non più velocemente di $(1/c)^i$ $(c > 1)$) assicura che $c_i\rho_k$ non sia troppo più piccolo di $c_{i-1}\rho_k$; infatti, se il passo ρ_k non va bene, si può sì ridurre il passo, ma per mantenere la convergenza del metodo tale riduzione deve essere la minima possibile (affinchè anche λ_k soddisfi una condizione del tipo $\lambda_k/\|h_x(\mathbf{x}_k)\| \geq \sigma' > 0$). È da notare che nella effettiva implementazione della tecnica (tp), si richiede spesso soltanto che c_j sia il primo termine della successione (succ) per cui

$$h(\mathbf{x}_k - c_j\rho_k\mathbf{s}_k) < h(\mathbf{x}_k)$$

Questa semplificazione della tecnica (tp) si può giustificare osservando che il numero χ, in teoria, può essere preso piccolo quanto si vuole.

Com'è evidente, sia la tecnica (tt) che quella (tp) (vedi i Teoremi 1 e 2) produconno una successione $\mathbf{x}_k$ convergente se imponiamo una limitazione superiore sulla scelta di ρ_k:

Corollario 1. *Arricchendo le ipotesi analitiche ed algoritmiche del Teorema 1 (oppure 2), rispettivamente, con l'ipotesi che in $\mathcal{I}_0$ non vi sia che un numero finito di punti stazionari per h e con l'ipotesi che*

$$\frac{\rho_k}{\|h_x(\mathbf{x}_k)\|} \leq \Sigma,$$

oltre alle conclusioni già note, potremo dire che $\mathbf{x}_k$ è convergente (converge a un punto stazionario per h).

Teorema 1 $\Rightarrow$ conv. dei metodi "lineari" del grad. e del rilass. ($\rho_k = \infty$)

Teorema 3 *Sia* $h(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b} + c$ *(*A $n \times n$ *reale simmetrica definita positiva,* $\mathbf{b} \in \mathbb{R}^n$, $c \in \mathbb{R}$*). Sia* $\mathbf{x}_0 \in \mathbb{R}^n$ *e, per* $k = 0, \ldots,$ *si ponga* $\mathbf{x}_{k+1} = \mathbf{x}_k - \lambda_k \mathbf{s}_k$ *con* $\mathbf{s}_k$ *e* λ_k *tali che*

$$\frac{h_x(\mathbf{x}_k)\mathbf{s}_k}{\|h_x(\mathbf{x}_k)\|} \geq \gamma > 0,$$

$$h(\mathbf{x}_k - \lambda_k \mathbf{s}_k) \leq h(\mathbf{x}_k - \lambda \mathbf{s}_k),\ \lambda \in \mathbb{R}^+, \qquad \textit{cioè } \lambda_k = \frac{\mathbf{s}_k^T h_x(\mathbf{x}_k)}{\mathbf{s}_k^T A \mathbf{s}_k}.$$

Allora $\{\mathbf{x}_k\}_{k=0}^{+\infty}$ *è convergente ad* $A^{-1}\mathbf{b}$ *(*$\forall \mathbf{x}_0$*).*

Dim. $\forall \mathbf{x}_0 \in \mathbb{R}^n$, $\{\mathbf{x} : h(\mathbf{x}) \leq h(\mathbf{x}_0)\}$ è compatto (connesso, convesso). Dal Teorema 1 per $\rho_k = +\infty$ segue che $\lim_{k\to\infty} h_x(\mathbf{x}_k) = \mathbf{0}$, cioè ogni punto di accumulazione di $\mathbf{x}_k$ è punto stazionario per h; ma h ha un solo punto stazionario, $A^{-1}\mathbf{b}$. Quindi $\mathbf{x}_k$ è convergente e converge a $A^{-1}\mathbf{b}$.

Applicazioni:

1) $\mathbf{s}_k = \frac{h_x(\mathbf{x}_k)}{\|h_x(\mathbf{x}_k)\|}$ (Grad.) soddisfa l'ipotesi con $\gamma = 1 \Rightarrow$ il metodo del Gradiente è convergente

2) $\mathbf{s}_k = \text{sign}\,(h_x(\mathbf{x}_k))_{i_k}\mathbf{e}_{i_k}$, con i_k tale che $|(h_x(\mathbf{x}_k))_{i_k}| \geq c|(h_x(\mathbf{x}_k))_j|$, $\forall j$, dove c è fissata $0 < c \leq 1$, (Ril.Gen.) soddisfa l'ipotesi con $\gamma = \frac{c}{\sqrt{n}}$:

$$\frac{h_x(\mathbf{x}_k)^T\mathbf{s}_k}{\|h_x(\mathbf{x}_k)\|} = \frac{\text{sign}\,(h_x(\mathbf{x}_k))_{i_k}(h_x(\mathbf{x}_k))_{i_k}}{\|h_x(\mathbf{x}_k)\|} = \frac{|(h_x(\mathbf{x}_k))_{i_k}|}{\|h_x(\mathbf{x}_k)\|} \geq c\frac{|(h_x(\mathbf{x}_k))_j|}{\|h_x(\mathbf{x}_k)\|},\ \forall j$$

da cui, quadrando e sommando su j, si ha:

$$\frac{h_x(\mathbf{x}_k)\mathbf{s}_k}{\|h_x(\mathbf{x}_k)\|} \geq \frac{c}{\sqrt{n}} > 0.$$

$\Rightarrow$ il metodo del Rilassamento Generalizzato è convergente. In particolare, si ha la convergenza del metodo del Rilassamento, dove i_k è scelto tale che

$$\frac{|(h_x(\mathbf{x}_k))_{i_k}|}{\sqrt{a_{i_k,i_k}}} \geq \frac{|(h_x(\mathbf{x}_k))_j|}{\sqrt{a_{jj}}},\ \forall j. \qquad (rilass)$$

---

metodo di Newton per $\mathbf{f} = \mathbf{0} \equiv$ metodo di Minimizzazione per $h = \mathbf{f}^T\mathbf{f}$

Si consideri il seguente sistema di n equazioni non lineari nelle n incognite x_1, ..., x_n,

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ f_2(x_1, \ldots, x_n) \\ \cdots \\ f_n(x_1, \ldots, x_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{bmatrix}$$

(si suppone che questo sistema abbia soluzione).

Se $\overline{\mathbf{x}}$ è una sua soluzione, cioè $\mathbf{f}(\overline{\mathbf{x}}) = \mathbf{0}$, allora $\overline{\mathbf{x}}$ è un punto di minimo (assoluto) per la funzione $h(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x}) = \sum f_i(x_1, \ldots, x_n)^2$ (in cui h vale zero). Quindi, $\overline{\mathbf{x}}$ è un punto stazionario per h, cioè risolve il seguente sistema, delle derivate parziali di h,

$$h_x(\mathbf{x}) = [\cdots 2\sum f_k(\mathbf{x})\frac{\partial f_k}{\partial x_i}\cdots] = 2\mathbf{f}(\mathbf{x})^T\mathbf{f}_x(\mathbf{x}) = \mathbf{0}^T.$$

(Si noti che $\mathbf{f}_x(\mathbf{x})$ è lo Jacobiano di $\mathbf{f}$).

Non vale il viceversa; infatti, $\mathbf{f}(\overline{\mathbf{x}})^T\mathbf{f}_x(\overline{\mathbf{x}}) = \mathbf{0}^T$ non implica $\mathbf{f}(\overline{\mathbf{x}}) = \mathbf{0}$ se $\det \mathbf{f}_x(\overline{\mathbf{x}}) = 0$.

Tuttavia, se riusciamo ad individuare una regione contenente un minimo assoluto $\overline{\mathbf{x}}$ per h (ovvero uno zero per $\mathbf{f}$) dove lo Jacobiano di $\mathbf{f}$ è non singolare, allora in questa regione i punti stazionari per h saranno punti di minimo assoluti per h, e, quindi, soluzioni del sistema non lineare dato.

È ben noto che per calcolare approssimazioni $\mathbf{x}_k$ sempre migliori degli zeri di $\mathbf{f}$ (ovvero, per risolvere numericamente il sistema non lineare dato) si può usare il seguente schema iterativo di Newton:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{f}_x(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k). \qquad (Newton)$$

Ovviamente, se la successione $\{\mathbf{x}_k\}$ così definita converge ad $\overline{\mathbf{x}} \mid \mathbf{f}(\overline{\mathbf{x}}) = \mathbf{0}$, contemporaneamente $\{\mathbf{x}_k\}$ convergerà ad un punto di minimo assoluto (in particolare stazionario) della funzione $h(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x})$. (Newton) può cioè essere visto come un metodo per la risoluzione del problema

$$\min_{\mathbf{x}\in\mathbb{R}^n} h(\mathbf{x}), \quad h(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x}).$$

Tuttavia, applicato così com'è, ad esempio a partire da $\mathbf{x}_0 \in \mathbb{R}^n$, non è detto che la successione $\{\mathbf{x}_k\}$ rimanga nell'insieme $\mathcal{I}_0 = \{\mathbf{x} : h(\mathbf{x}) \le h(\mathbf{x}_0)\}$, perchè la disuguaglianza

$$h(\mathbf{x}_{k+1}) < h(\mathbf{x}_k) \qquad (decr)$$

può non essere soddisfatta. Notiamo che senza la condizione $\{\mathbf{x}_k\} \subset \mathcal{I}_0$, la successione $\{\mathbf{x}_k\}$ potrebbe non avere punti di accumulazione; in tal caso, la convergenza a zero di $\|h_x(\mathbf{x}_k)\|$ potrebbe verificarsi solo se $\mathbf{x}_k \to +\infty$.

Osserviamo che

$$h_x(\mathbf{x}_k)(-\mathbf{f}_x(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)) < 0$$

a meno che $\mathbf{f}(\mathbf{x}_k) = \mathbf{0}$. Quindi, $-\mathbf{f}_x(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)$, il passo usato in (Newton), è una direzione di decrescita per la funzione h; ne segue che possiamo dotare il metodo di Newton della condizione (decr) di decrescita di h, per ogni iterazione k, *modificando* (Newton) come segue:

$$\begin{array}{l}\text{individua il max } c,\ c \le 1, \ \text{ tale che} \\ h(\mathbf{x}_k - c\mathbf{f}_x(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)) < h(\mathbf{x}_k);\ \text{poni } \mathbf{x}_{k+1} = \mathbf{x}_k - c\mathbf{f}_x(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k).\end{array} \qquad (NewMod)$$

Il metodo di minimizzazione per la funzione $h(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x})$ appena abbozzato in (NewMod) può essere definito più rigorosamente e, allo stesso tempo, corredato di un teorema di convergenza globale, se inquadrato nella tecnica di minimizzazione (tp).

Innanzitutto, si riscrive il metodo di Newton (Newton) nella forma

$$\begin{array}{l}\mathbf{x}_{k+1} = \mathbf{x}_k - \rho_k\mathbf{s}_k \\ \rho_k = \|\mathbf{f}_x(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k)\|, \quad \mathbf{s}_k = \frac{1}{\rho_k}\mathbf{f}_x(\mathbf{x}_k)^{-1}\mathbf{f}(\mathbf{x}_k).\end{array}$$

Quindi, posto $h(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x})$, si osserva che $\rho_k h_x(\mathbf{x}_k)\mathbf{s}_k = 2h(\mathbf{x}_k)$, che valgono le disuguaglianze

$$\frac{h_x(\mathbf{x}_k)\mathbf{s}_k}{\|h_x(\mathbf{x}_k)\|} \ge \frac{1}{\|\mathbf{f}_x(\mathbf{x}_k)\|\|\mathbf{f}_x(\mathbf{x}_k)^{-1}\|}, \ \frac{\rho_k}{\|h_x(\mathbf{x}_k)\|} \ge \frac{1}{2\|\mathbf{f}_x(\mathbf{x}_k)\|^2}, \qquad (DIS)$$

e che l'algoritmo suggerito dalla tecnica (tp), corrispondente alle nostre h, ρ_k e $\mathbf{s}_k$, diventa:

$$\begin{array}{l} \mathbf{x}_{k+1} = \mathbf{x}_k - c_j \rho_k \mathbf{s}_k \text{ dove} \\ h(\mathbf{x}_k - c_j \rho_k \mathbf{s}_k) < (1 - 2\chi c_j) h(\mathbf{x}_k) \text{ e} \\ h(\mathbf{x}_k - c_i \rho_k \mathbf{s}_k) \geq (1 - 2\chi c_i) h(\mathbf{x}_k),\ i = 0, \ldots, j-1 \end{array} \qquad (NewModP)$$

(Nota: χ va scelto minore di $\frac{1}{2}$. Inoltre, la condizione su c_j in (NewModP) viene sostituita spesso con la condizione $h(\mathbf{x}_k - c_j \rho_k \mathbf{s}_k) < h(\mathbf{x}_k)$)

Utilizzando le disuguaglianze (DIS) e il Teorema 2 (tp), è semplice dimostrare il seguente risultato di convergenza per l'algoritmo (NewModP):

Teorema 3. *Sia $\mathbf{f} \in C^1$ e con Jacobiano $\mathbf{f}_x$ non singolare nell'insieme compatto $\mathcal{I}_0 = \{\mathbf{x} : h(\mathbf{x}) \leq h(\mathbf{x}_0)\}$, $h = \mathbf{f}^T\mathbf{f}$. Se il sistema $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ è risolvibile, allora il metodo di Neton modificato (NewModP), inizializzato con $\mathbf{x}_0$, converge a una sua soluzione.*

Commento (più chiaro con l'aiuto di disegni, considerando il caso unidimensionale). L'ipotesi fondamentale utilizzata nel teorema di convergenza è che lo Jacobiano di $\mathbf{f}$ sia non singolare nell'insieme compatto $\mathcal{I}_0$. Se tale ipotesi non è soddisfatta, come succede in pratica, h può avere, in $\mathcal{I}_0$, punti stazionari (e.g. minimi locali) che non sono minimi assoluti e, quindi, non risolvono il sistema non lineare $\mathbf{f}(\mathbf{x}) = \mathbf{0}$. Inoltre, in tal caso, non esiste un risultato teorico che assicuri la convergenza della successione $\{\mathbf{x}_k\}$. Il metodo di Newton mo dificato, infatti, potrebbe convergere ad un punto stazionario non minimo assoluto, al minimo assoluto, o non convergere affatto. In generale, non è mai noto a priori che in $\mathcal{I}_0$ lo Jacobiano di $\mathbf{f}$ sia non singolare. L'utente si accorge che esso è singolare in qualche punto di $\mathcal{I}_0$ quando, durante l'applicazione di (NewModP), j è scelto dall'algoritmo $\neq 0$ per diversi valori di k, oppure quando $\mathbf{x}_k$ sembra convergere, ma $h(\mathbf{x}_k)$ rimane grande. In questi casi, conviene ripartire da un'altra zona (ridefinire $\mathbf{x}_0$), perchè o non si sta convergendo o si sta convergendo a un minimo locale. Durante l'applicazione di (NewModP) l'utente dovrebbe anche verificare che le componenti $f_i(\mathbf{x})^2$ della funzione $h(\mathbf{x})$ abbiano tutte più o meno gli stessi ordini di grandezza e che decrescano contemporaneamente. Perchè avvenga ciò è a volte necessario applicare degli opportuni pesi p_i alle f_i^2 e continuare il procedimento di minimizzazione sulla funzione $\sum p_i f_i(\mathbf{x})^2$.

Le dimostrazioni dei Teoremi 1, 2 e del Corollario 1

Dim. del Teorema 1

La funzione unidimensionale $\varphi_k(\lambda) = h(\mathbf{x}_k - \lambda \mathbf{s}_k)$ in zero è decrescente, infatti $\varphi_k'(0) = -h_x(\mathbf{x}_k)\mathbf{s}_k < 0$. Quindi,

$$h(\mathbf{x}_{k+1}) = \varphi_k(\lambda_k) = \min_{0 < \lambda \leq \rho_k} \varphi_k(\lambda) < \varphi_k(0) = h(\mathbf{x}_k)$$

e $\mathbf{x}_k \in \mathcal{I}_0$, $\forall k$.

Sia, per assurdo, $\overline{\mathbf{x}} \in \mathcal{I}_0$ un punto di accumulazione per $\mathbf{x}_k$ tale che $h_x(\overline{\mathbf{x}}) \neq 0$ e $\overline{\mathbf{x}}_l$ una sottosuccessione di $\mathbf{x}_k$ convergente a $\overline{\mathbf{x}}$. Notiamo che se $\overline{\mathbf{x}}_l = \mathbf{x}_k$, allora $\overline{\mathbf{x}}_{l+1} = \mathbf{x}_m$, $m \geq k+1$, e $h(\overline{\mathbf{x}}_{l+1}) \leq h(\mathbf{x}_{k+1})$. Inoltre, esiste una sfera chiusa S_δ di centro $\overline{\mathbf{x}}$ e raggio $\delta > 0$ in cui $\|h_x(\mathbf{x})\| \geq M = \frac{1}{2}\|h_x(\overline{\mathbf{x}})\| > 0$ ed un intero positivo N tale che, per $l > N$, $\|\overline{\mathbf{x}}_l - \overline{\mathbf{x}}\| < \delta/2$. Sia $\mu = \mu(M, \chi, \gamma)$ il numero positivo del Lemma 1 applicato a $\mathcal{K} = S_\delta$. Sia l fissato $> N$ e k tale che $\overline{\mathbf{x}}_l = \mathbf{x}_k$ e si consideri il punto di S_δ

$$\mathbf{x}_k - \Lambda \mathbf{s}_k,\ \Lambda = \min\{\mu, \frac{\delta}{2}, \frac{1}{2}\sigma\|h_x(\overline{\mathbf{x}})\|\}.$$

(Esercizio: perchè è in S_δ ?)
Poiché $\Lambda \le \rho_k$

(Esercizio: verificarlo)
e $\Lambda \le \mu$, si ha

$$h(\overline{\mathbf{x}}_{l+1}) \le h(\mathbf{x}_{k+1}) \le h(\mathbf{x}_k - \Lambda \mathbf{s}_k) < h(\overline{\mathbf{x}}_l) - \frac{1}{2}\Lambda\chi\gamma\|h_x(\overline{\mathbf{x}})\|.$$

Passando al limite per $l \to \infty$, si ha l'assurdo $\|h_x(\overline{\mathbf{x}})\| < 0$. Quindi non può essere $\|h_x(\overline{\mathbf{x}})\| \neq \mathbf{0}$.

Dim. del Teorema 2

Per ogni k tale che $h_x(\mathbf{x}_k) \neq \mathbf{0}$ esiste una sfera chiusa $\mathcal{K}$ di centro $\mathbf{x}_k$ dove $\|h_x(\mathbf{x})\| \ge M_k > 0$. Quindi, per il Lemma 1, applicato a $\mathcal{K}$, esiste $\mu = \mu(M_k, \chi, \gamma) > 0$ tale che

$$\lambda \le \mu \text{ e } \mathbf{x}_k - \lambda \mathbf{s}_k \in \mathcal{K} \;\Rightarrow\; h(\mathbf{x}_k - \lambda \mathbf{s}_k) < h(\mathbf{x}_k) - \lambda\chi h_x(\mathbf{x}_k)\mathbf{s}_k.$$

Ne segue l'esistenza dell'indice j usato in (*) nella definizione del passo λ_k. Quindi, ogni iterazione dell'algoritmo è ben definita (finché $h_x(\mathbf{x}_k) \neq \mathbf{0}$).

Per costruzione si ha immediatamente che $h(\mathbf{x}_{k+1}) < h(\mathbf{x}_k)$ e $\mathbf{x}_{k+1} \in \mathcal{I}_0$ per ogni $k \ge 0$.

Sia, per assurdo, $\overline{\mathbf{x}} \in \mathcal{I}_0$ un punto di accumulazione per $\mathbf{x}_k$ tale che $h_x(\overline{\mathbf{x}}) \neq 0$ e $\overline{\mathbf{x}}_l$ una sottosuccessione di $\mathbf{x}_k$ convergente a $\overline{\mathbf{x}}$. Notiamo che se $\overline{\mathbf{x}}_l = \mathbf{x}_k$, allora $\overline{\mathbf{x}}_{l+1} = \mathbf{x}_m$, $m \ge k+1$, e $h(\overline{\mathbf{x}}_{l+1}) \le h(\mathbf{x}_{k+1})$. Inoltre, esiste una sfera chiusa S_δ di centro $\overline{\mathbf{x}}$ e raggio $\delta > 0$ in cui $\|h_x(\mathbf{x})\| \ge M = \frac{1}{2}\|h_x(\overline{\mathbf{x}})\| > 0$ ed un intero positivo N tale che, per $l > N$, $\|\overline{\mathbf{x}}_l - \overline{\mathbf{x}}\| < \delta$. Sia $\mu = \mu(M, \chi, \gamma)$ il numero positivo del Lemma 1 applicato a $\mathcal{K} = S_\delta$. Sia l fissato $> N$ e k tale che $\overline{\mathbf{x}}_l = \mathbf{x}_k$. Si osserva che

$$\lambda_k > \frac{1}{c}\Lambda,\ \Lambda = \min\{\mu, \frac{1}{2}\sigma\|h_x(\overline{\mathbf{x}})\|\}$$

($\lambda_k = \rho_k \Rightarrow \lambda_k \ge \Lambda > \Lambda/c$; $\lambda_k = c_j\rho_k \Rightarrow c_{j-1}\rho_k > \mu \Rightarrow \lambda_k > c_j\mu/c_{j-1} \ge \Lambda/c$). Quindi

$$h(\overline{\mathbf{x}}_{l+1}) \le h(\mathbf{x}_{k+1}) < h(\mathbf{x}_k) - \lambda_k\chi\gamma\|h_x(\mathbf{x}_k)\| \le h(\overline{\mathbf{x}}_l) - \frac{1}{c}\Lambda\chi\gamma\frac{1}{2}\|h_x(\overline{\mathbf{x}})\|.$$

La tesi segue come per il Teorema 1.

Dim. del Corollario 1

Racchiudiamo in sfere C_γ aperte disgiunte i punti stazionari di h in $\mathcal{I}_0$. Sia $d > 0$ la minima distanza tra tali sfere; quindi, se $\mathbf{x} \in C_\gamma$, $\tilde{\mathbf{x}} \in C_\beta$ e $\gamma \neq \beta$, allora $\|\mathbf{x} - \tilde{\mathbf{x}}\| > d$. Inoltre, sia $\mu = \min_{\mathcal{I}_0 \setminus \cup C_g a} \|h_x(\mathbf{x})\|$. Si noti che $\mu > 0$. Poiché $h_x(\mathbf{x}_k) \to \mathbf{0}$ (perchè?), esiste un intero positivo N tale che, per $k > N$, $\|h_x(\mathbf{x}_k)\| < \min\{\mu, d/\Sigma\}$. In particolare, gli elementi con $k > N$ della successione $\{\mathbf{x}_k\}$ appartengono all'insieme $\cup C_\gamma$. In realtà si può affermare che tali elementi devono essere tutti contenuti nella stessa sfera, diciamo C_γ, in cui è contenuto $\mathbf{x}_{N+1}$.

(Esercizio: dimostrare che se $\mathbf{x}_k \in C_\gamma$, $k > N$, allora $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| < d$)

Da ciò segue la convergenza della successione $\{\mathbf{x}_k\}$ al punto stazionario racchiuso in C_γ (se non convergesse, o, equivalentemente, se avesse in C_γ due o più punti di accumulazione, questi sarebbero punti stazionari per h, il che è assurdo perchè C_γ contiene solo un punto stazionario).

mettere qui la cosa che si vuole stampare ...