18. DISCRETIZATION AND NUMERICAL STABILITY

We leave the mathematicians' ideal world of real and complex numbers to see how the algorithms considererd in the last section can be implemented on a computer. We shall also estimate the effects of numerical errors.

Computer arithmetic

It is not possible to represent an arbitrary real number on a computer. Given a machine base β, precision t, underflow limit L, and overflow limit U, we can represent only the numbers

$$\pm \cdot d_1 \ldots d_t \times \beta^e , \quad 0 \leq d_i < \beta , \quad d_1 \neq 0 , \quad L \leq e \leq U ,$$

together with the number 0. These are known as the floating-point numbers. The value of (β, t, L, U) for Cyber 180 Model 840 is (2, 48, -4096, 4095), while for Cray-1 it is (2, 48, -16384, 8191). An arbitrary real number is 'approximately represented' by its nearest floating-point neighbour if rounded arithmetic is used; in case of a tie, it is rounded away from zero. A complex number is represented by the pair of floating-point representations of its real and imaginary parts. The errors introduced by this approximate representation while performing the arithmetic operations +, −, ×, / are known as the round-off errors. One of the ways of reducing these errors is to carry out certain operations in higher precision, like double (2t) precision or extended (4t) precision. Taking the inner product

$$(18.1) \qquad \underset{\sim}{y}^H \underset{\sim}{x} = x(1)\overline{y(1)} + \ldots + x(n)\overline{y(n)}$$

of two n-vectors $\underset{\sim}{x}$ and $\underset{\sim}{y}$ is one such operation. In the multiplication of two single precision numbers, some computers calculate the entire 2t-digit product. If the additions in (18.1) are also performed in 2t precision, then we obtain the so called *double precision accumulation of inner products.* (See [ST], p.73.) This is particularly useful in calculating the residual $\underset{\sim}{r}_0 = A\underset{\sim}{x}_0 - \underset{\sim}{b}$ in double precision, where $\underset{\sim}{x}_0$ is an approximate initial solution of the linear system $A\underset{\sim}{x} = \underset{\sim}{b}$; in this case $\underset{\sim}{x}_1 = \underset{\sim}{x}_0 - A^{-1}\underset{\sim}{r}_0$ turns out to be a refinement of x_0 . (Cf. Problem 5.1.) A similar remark holds for the iterative refinement of eigenelements. (Cf. Problem 11.7.) (See [ST] Algorithms 4.5.1 and 5.4.1; [FM], pp.49-54.)

Discretization

If a Banach space X is finite dimensional, then it can be identified with $\mathbb{C}^M$, where M is the dimension of X . Also, $x \in X$ can be represented by a column vector $\underset{\sim}{x}$ with M complex entries. If X is infinite dimensional, we consider a sequence (π_n) of projections in $BL(X)$ such that $\pi_n x \to x$ for every $x \in X$, and for a large enough positive integer M , approximate x by $\pi_M x$. Then there are $f_1,\ldots,f_M$ in X and $f_1^*,\ldots,f_M^*$ in X^* such that

$$\langle f_j, f_i^* \rangle = \delta_{i,j} , \quad i,j = 1,\ldots,M ,$$
$$\pi_M x = \langle x, f_1^* \rangle f_1 + \ldots + \langle x, f_M^* \rangle f_M \quad \text{for all} \quad x \in X .$$

We say that $x \in X$ is *discretized* by the column vector

$$\underset{\sim}{x} = [\langle x, f_1^* \rangle, \ldots, \langle x, f_M^* \rangle]^t .$$

Let $T \in BL(X)$ and $x \in X$. Then Tx is discretized by the column vector

$$[\langle Tx, f_1^* \rangle, \ldots, \langle Tx, f_M^* \rangle]^t .$$

Often it is not possible to find $\langle Tx, f_i^* \rangle$ $(i = 1,\ldots,M)$ exactly, in which case we approximate it by $\langle T\pi_M x, f_i^* \rangle$. Then Tx is (approximately) discretized by the column vector

$$[TM]\underset{\sim}{x} ,$$

where

$$[TM] = [\langle Tf_j, f_i^* \rangle] , \quad i,j = 1,\ldots,M .$$

Thus, essentially we replace T by $\pi_M T \pi_M$. If even the scalar products $\langle Tf_j, f_i^* \rangle$, $i,j = 1,\ldots,M$, cannot be calculated exactly, we consider a close approximation $\tilde{T}$ of the operator T and calculate $\langle \tilde{T}f_j, f_i^* \rangle$ instead. For example, let $X = C([a,b])$ and T be the integral operator

$$Tx(s) = \int_a^b k(s,t)x(t)dt , \quad x \in X , \quad s \in [a,b] ,$$

where the kernel k is continuous. Then we can consider

$$\tilde{T}x(s) = \sum_{j=1}^{M} w_j^{(M)} k(s, t_j^{(M)}) x(t_j^{(M)}) ,$$

where the nodes $t_1^{(M)},\ldots,t_M^{(M)}$ in $[a,b]$ and the weights $w_1^{(M)},\ldots,w_M^{(M)}$ in $\mathbb{C}$ give a convergent quadrature formula (cf. (16.5))

$$Qx = \sum_{j=1}^{M} w_j^{(M)} x(t_j^{(M)}) , \quad x \in X .$$

Coming now to the algorithms given in the last section, we observe that they depend on the choice of a finite rank operator $T_0 \in BL(X)$:

$$T_0 x = \sum_{i=1}^{n} \langle x, x_i^* \rangle x_i , \quad x \in X ,$$

with $x_1,\ldots,x_n$ in X and $x_1^*,\ldots,x_n^*$ in X^*.

In the first step of the algorithms we solve an eigenvalue problem for the matrix

$$A = [\langle x_j, x_i^* \rangle] , \quad i,j = 1,\dots,n .$$

We look for a nonzero simple eigenvalue λ_0 of A. Depending on which eigenvalue λ of T we wish to approximate, we choose such a λ_0 and find a corresponding eigenvector $\underset{\sim}{u} \in \mathbb{C}^n$ of A. In order to economize computer memory and time, we choose n relatively small; in fact, much smaller than M, say, $n = 10$, if $M = 100$. We then find an eigenvector $\underset{\sim}{v} \in \mathbb{C}^n$ of

$$A^H = [\langle x_j^*, x_i \rangle] , \quad i,j = 1,\dots,n ,$$

corresponding to $\bar{\lambda}_0$, which satisfies

$$\underset{\sim}{v}^H \underset{\sim}{u} = 1/\lambda_0 .$$

In the second step, we calculate for $j = 1,2,\dots,$ the j-th eigenvalue iterate λ_j and the j-th eigenvector iterate φ_j. To start with, we let

$$\varphi_0 = u(1)x_1 + \dots + u(n)x_n .$$

Since

$$\underset{\sim}{x}_i = [\langle x_i, f_i^* \rangle, \dots, \langle x_i, f_M^* \rangle]^t , \quad i = 1,\dots,n ,$$

we see that φ_0 is discretized by

$$\underset{\sim}{c}_0 = [AV]\underset{\sim}{u} ,$$

where

$$[AV] = [\langle x_j, f_i^* \rangle] , \quad i = 1,\dots,M , \quad j = 1,\dots,n .$$

This matrix is vertical in the sense that it has many more rows than columns; it is, so to say, the matrix $A = [\langle x_j, x_i^* \rangle]$ made vertical. Hence the name $[AV]$.

To calculate the eigenvalue iterates

$$\lambda_j = \langle T\varphi_{j-1}, \varphi_0^* \rangle = \sum_{i=1}^{n} \langle T\varphi_{j-1}, x_i^* \rangle \overline{v(i)} ,$$

we must find $\langle Tx, x_i^* \rangle$, $i = 1,\dots,n$, for various $x \in X$. Replacing x by $\pi_M x = \sum_{j=1}^{M} \langle x, f_j^* \rangle f_j$ we see that

$$[\langle T\pi_M x, x_1^* \rangle, \dots, \langle T\pi_M x, x_n^* \rangle]^t = [TAH]\underset{\sim}{x} ,$$

where

$$[TAH] = [\langle Tf_j, x_i^* \rangle] , \quad i = 1,\dots,n , \quad j = 1,\dots,M .$$

This matrix is horizontal in the sense that it has many more columns than rows; hence the letter H in its name. For the algorithms 17.10 and 17.11, we also need to find

$$\mu_j = \sum_{i=1}^{n} \langle T^2\varphi_{j-1}, x_i^* \rangle \overline{v(i)} .$$

This can be done by noting that

$$[\langle T^2\pi_M x, x_i^* \rangle, \dots, \langle T^2\pi_M x, x_n^* \rangle]^t = [T2AH]\underset{\sim}{x} ,$$

where

$$[T2AH] = [\langle T^2 f_j, x_i^* \rangle] , \quad i = 1,\dots,n , \quad j = 1,\dots,M .$$

Thus, if φ_{j-1} is discretized by $\underset{\sim}{c}_{j-1} \in \mathbb{C}^M$, then

$$\lambda_j = \underset{\sim}{v}^H [TAH] \underset{\sim}{c}_{j-1} \quad \text{and} \quad \mu_j = \underset{\sim}{v}^H [T2AH] \underset{\sim}{c}_{j-1} .$$

To calculate the eigenvector iterate φ_j , we need to solve a system

$$\underset{\sim}{v}^H \underset{\sim}{\alpha} = 0 , \quad (A-\lambda_0 I)\underset{\sim}{\alpha} = \underset{\sim}{\beta}_j ,$$

of $n + 1$ equations in the n unknowns $\alpha(1),\dots,\alpha(n)$; the given vector $\underset{\sim}{\beta}_j$ satisfies $\underset{\sim}{v}^H \underset{\sim}{\beta}_j = 0$, and is determined by the information available at this stage. Let us denote the solution by

$\underset{\sim}{\alpha}_j = [\alpha_j(1),\ldots,\alpha_j(n)]^t$.

The iterate φ_j is a linear combination of $x_1,\ldots,x_n$, $\varphi_0,\ldots,\varphi_{j-1}$, $T\varphi_{j-1}$ and $T^2\varphi_{j-1}$, whose coefficients are determined by $\lambda_0,\ldots,\lambda_j$, μ_j , $\alpha_j(1),\ldots,\alpha_j(n)$. The matrices [AV], [TM] and

$$[T2M] = [\langle T^2 f_j, f_i^* \rangle] \ , \quad i,j = 1,\ldots,M \ ,$$

allow us to discretize φ_j , $j = 1,2,\ldots$. For example, in Algorithm 17.9 for the fixed point iteration scheme we have

$$\varphi_j = \frac{1}{\lambda_0}\left[\alpha_j(1)x_1+\ldots+\alpha_j(n)x_n + (\lambda_0-\lambda_j)\varphi_{j-1} + T\varphi_{j-1}\right] .$$

It is discretized by

$$\underset{\sim}{c}_j = \frac{1}{\lambda_0}\left[[AV]\underset{\sim}{\alpha}_j + (\lambda_0-\lambda_j)\underset{\sim}{c}_{j-1} + [TM]\underset{\sim}{c}_{j-1}\right] .$$

Accuracy of the approximations

In the case of all the algorithms of Section 17, the eigenvalue and eigenvector iterates λ_j and φ_j converge, under suitable conditions (given in Section 14), to a nonzero simple eigenvalue λ and a corresponding eigenvector φ of T , which satisfy $\langle T\varphi,\varphi_0^*\rangle = \lambda\langle\varphi,\varphi_0^*\rangle = \lambda$, where $\varphi_0^* = v(1)x_1^* + \ldots + v(n)x_n^*$. For the Rayleigh-Schrödinger scheme (11.18) and the fixed point scheme (11.19), this eigenvalue λ of T is the nearest spectral point of T to λ_0 . (cf. Theorem 11.8.) Since we have replaced the possibly infinite dimensional operator T on X by the M-dimensional operator [TM] , and $x \in X$ by $\pi_M x$ in the discretization procedure, the computed λ_j's will converge to a nonzero simple eigenvalue $\lambda^{(M)}$ of [TM] , and the discretizations $\underset{\sim}{c}_j$ of φ_j will converge to the corresponding eigenvector $\underset{\sim}{c}^{(M)}$ of [TM] which satisfies

$$\underset{\sim}{v}^H[TAH]\underset{\sim}{c}^{(M)} = \lambda^{(M)} .$$

The whole point of going through the iterations is to approximate the nonzero simple eigenvalues of [TM] (or T) as closely as we wish without solving the large (or infinite dimensional) eigenvalue problem for [TM] (or T) . However, for illustrative purposes we may compute the eigenelements $\lambda^{(M)}$ and $\underset{\sim}{c}^{(M)}$ of [TM] directly to get an idea of the actual accuracy reached at the j-th iteration, $j = 1,2,\ldots$.

All the same, we must have criteria for deciding when a sufficient accuracy is reached without actually knowing $\lambda^{(M)}$ and $\varphi^{(M)}$. If such criteria are satisfied, the iteration should be stopped. The degree of accuracy can be measured either by the norm of the <u>residual</u>

$$r_j = T\varphi_{j-1} - \lambda_j\varphi_{j-1} , \quad j = 1,2,\ldots,$$

or by the <u>relative increment</u>

$$d_j = \|\varphi_j-\varphi_{j-1}\| \,/\, \|\varphi_j\|$$

between two successive iterates. Note that if φ_{j-1} is actually an eigenvector of T , then since $\langle\varphi_{j-1},\varphi_0^*\rangle = 1$, we see that $\lambda_j = \langle T\varphi_{j-1},\varphi_0^*\rangle$ must be the corresponding eigenvalue of T , so that $r_j = 0$. If X is a Hilbert space, then one can compute the Rayleigh quotient

$$q(\varphi_{j-1}) = \langle T\varphi_{j-1},\varphi_{j-1}\rangle/\langle\varphi_{j-1},\varphi_{j-1}\rangle$$

and the corresponding residual

$$r_j' = \|T\varphi_{j-1}-q(\varphi_{j-1})\varphi_{j-1}\|_2$$

in view of the minimum residual property (8.9) of $q(\varphi_{j-1})$.

Thus, we stop the iteration if

$$\|r_j\| < \beta^{-t_0} \text{ and/or } d_j < \beta^{-t_0} ,$$

where β is the machine base, and t_0 is a given positive integer less than or equal to the machine precision t. In order to avoid being on the boundary of the precision, we take $t_0 = t - 1$ or $t - 2$. If $X = C([0,1])$ with the sup norm, then we can employ the $\| \|_\infty$ norm on $\mathbb{C}^M$, and if X is a Hilbert space, we can use the Euclidean norm $\| \|_2$ on $\mathbb{C}^M$.

There is usually a trade-off between the size n of the matrix A of the initial eigenvalue problem and the number of iterations needed to attain a desired accuracy. It is economical to choose a smaller n and opt for a greater number of iterations. The example in Table 19.9 will illustrates this point.

If one needs a highly accurate eigenvalue approximation but only a moderately accurate eigenvector approximation, then it is perhaps more practical to carry out two iteration processes simultaneously: one on the eigenpair (λ_0,φ_0) of T_0 and the other on the eigenpair $(\bar{\lambda}_0,\varphi_0^*)$ of T_0^*, as pointed out in Remark 11.9(iv). The generalized Rayleigh quotient based at (φ_j,φ_j^*), namely

$$q_j = \langle T\varphi_j,\varphi_j^*\rangle / \langle \varphi_j,\varphi_j^*\rangle$$

will then be an approximation of the eigenvalue λ of T of a much higher order, provided $\|(T^*-T_0^*)\varphi_0^*\|$ and $\|(T^*-T_0^*)S_0^*\|$ are small. (Cf. (11.28).) If T and T_0 are self-adjoint, then it is easy to compute q_j since $\varphi_j^* = \varphi_j$, so that we do not need to carry out two iteration processes. In this case, for the Rayleigh-Schrödinger iteration scheme (11.18), the eigenvalue iterates λ_{2j} and λ_{2j+1} can also be computed on knowing φ_j (cf. (10.9)).

Finally, we make some comments on the relative merits of the algorithms for the four iterations (11.18), (11.19), (11.31) and (11.35). The Rayleigh-Schrödinger iteration involves computing the sum

$$\lambda_1\varphi_{j-1} + (\lambda_2-\lambda_1)\varphi_{j-2} + \dots + (\lambda_j-\lambda_{j-1})\varphi_0$$

at the j-th step, where the coefficients $(\lambda_j-\lambda_{j-1})$ become progressively small as j increases. This is undesirable in a floating-point arithmetic. From this point of view, the fixed point iteration (11.31) should be preferred. The modified fixed point iteration (11.31) and the Ahués iteration (11.35) involve additional computations of $T^2\varphi_{j-1}$ and $\langle T^2\varphi_{j-1},\varphi_0^*\rangle$ at the j-th stage and as such, are more expensive than the fixed point iteration (11.18). However, the numerical experiments given in Tables 19.3, 19.4 and 19.5 indicate that the iterations (11.31) and (11.35) give the desired accuracy very fast; the iteration (11.31) often converges faster than the iteration (11.35), which was regarded as the best among those considered in [A]. (See p.157 of [A].)

Numerical stability

The round-off errors caused by floating-point arithmetic can sometimes assume alarming proportions. A minor change in the data can give rise to a major deviation in the solution of an eigenvalue problem, or of a system of linear equations. If we employ a 'DO' loop in an iteration process, the errors can accumulate and cause an overflow or an underflow. It is advisable, then, to consider some conditions which, when satisfied, preclude the possibility of small errors in the initial stage leading up to large errors in the final stage. In that case, the computations are said to be numerically stable.

While implementing the algorithms of Section 17, questions of numerical stability arise at three places: (i) calculation of the eigenelements λ_0 and $\underset{\sim}{u}$ of A, (ii) calculation of the eigenvector $\underset{\sim}{v}$ of A^* corresponding to $\bar{\lambda}_0$, and (iii) solution of the system

$$\underset{\sim}{v}^H\underset{\sim}{x} = 0 \ , \quad (A-\lambda_0 I)\underset{\sim}{x} = \underset{\sim}{\beta}$$

of linear equations. We shall now take up these questions one by one.

Initial eigenvalue problem for A

The entries $\langle x_j, x_i^* \rangle$, $i,j = 1,\ldots,n$, of the matrix A can, in general, be calculated only approximately. For some important eigenvalue routines, the computed eigenvalues are in fact, the eigenvalues of a nearby matrix $\hat{A}$ ([GV], p.200). Thus, instead of finding $0 \neq \lambda_0 \in \mathbb{C}$ and $\underset{\sim}{0} \neq \underset{\sim}{u} \in \mathbb{C}^n$ such that

$$A\underset{\sim}{u} = \lambda_0 \underset{\sim}{u} \ ,$$

we actually find $0 \neq \hat{\lambda}_0 \in \mathbb{C}$ and $\underset{\sim}{0} \neq \hat{\underset{\sim}{u}} \in \mathbb{C}^n$ such that

$$\hat{A}\hat{\underset{\sim}{u}} = \hat{\lambda}_0 \hat{\underset{\sim}{u}} \ .$$

Let $E = \hat{A} - A$ denote the error matrix as well as the induced operator on $\mathbb{C}^n$. Let $\mathcal{P}$ and $\mathcal{S}$ denote the spectral projection and the reduced resolvent associated with A and λ_0, respectively.

Let $\| \ \|$ denote a norm on $\mathbb{C}^n$, and let $\| \ \|_*$ be the induced norm on the adjoint space. For example, if $\| \ \| = \| \ \|_p$, then $\| \ \|_* = \| \ \|_q$, where $1/p + 1/q = 1$, $1 \leq p \leq \infty$ ..

Assume that A has a simple nonzero eigenvalue λ_0 and a corresponding eigenvector $\underset{\sim}{u}$, and let $\underset{\sim}{w}$ denote the eigenvector of A^* corresponding to $\bar{\lambda}_0$ such that $\underset{\sim}{w}^H\underset{\sim}{u} = 1$. (See Theorem 8.3.)

Assume that

$$0 < \gamma_0 = \max\{\|E\underset{\sim}{u}\|\ \|\underset{\sim}{w}\|_*\|\mathcal{S}\|\ ,\ \|E\mathcal{S}\|\} < 1/4 .$$

By applying Theorem 11.5 to $T_0 = A$ and $T = \hat{A}$, we see that $\hat{A}$ has an eigenvalue $\hat{\lambda}_0$ and a corresponding eigenvector $\hat{\underset{\sim}{u}}$ satisfying $\underset{\sim}{w}^H\hat{\underset{\sim}{u}} = 1$ such that

$$\|\hat{\underset{\sim}{u}}-\underset{\sim}{u}\| \leq \|E\underset{\sim}{u}\|\ \|\mathcal{S}\|\ \frac{g(\gamma_0) - 1}{\gamma_0} ,$$

$$|\hat{\lambda}_0-\lambda_0| \leq \|E\underset{\sim}{u}\|\ \|\underset{\sim}{w}\|_* g(\gamma_0) ,$$

where $g(t) = 1 + t + a_2t^2 + \ldots$ is the function defined by (11.1) for $|t| \leq 1/4$. (See (11.20) and (11.21).) Also, Theorem 11.8 can be employed to conclude that $\hat{\lambda}_0$ is a simple eigenvalue of $\hat{A}$. It will be nonzero if it is sufficiently near λ_0 .

Now, assume that the eigenvector $\underset{\sim}{u}$ of A is scaled so that $\|\underset{\sim}{u}\| = 1$. Then

$$\mathcal{P}\underset{\sim}{x} = (\underset{\sim}{w}^H\underset{\sim}{x})\underset{\sim}{u} ,\quad \|\mathcal{P}\| = \|\underset{\sim}{w}\|_* .$$

Also,

$$|(g(t)-1)/t| = 1 + O(|t|) = |g(t)| \quad \text{as} \quad |t| \to 0 ,$$

and

$$\gamma_0 \leq \|E\|\ \|\mathcal{P}\|\ \|\mathcal{S}\| .$$

Hence

(18.1) $$\|\hat{\underset{\sim}{u}}-\underset{\sim}{u}\| \leq \|E\|\ \|\mathcal{S}\| + O(\|E\|^2) ,$$

(18.2) $$|\hat{\lambda}_0-\lambda_0| \leq \|E\|\ \|\mathcal{P}\| + O(\|E\|^2) .$$

Thus, a small error of size $\|E\|$ in the formation of the matrix A can cause an error of size at most $\|E\|\ \|\mathcal{P}\|$ in the eigenvalue λ_0 of A and of size at most $\|E\|\ \|\mathcal{S}\|$ in a corresponding eigenvector $\underset{\sim}{u}$ of norm 1 .

For this reason, $\|\mathscr{P}\|$ is called the condition number for finding the simple eigenvalue λ_0, and $\|\mathscr{S}\|$ is called the condition number for finding the corresponding unit eigenvector $\underset{\sim}{u}$ of A. If a condition number is small, the relevant problem is said to be well-conditioned, otherwise it is ill-conditioned. Here are some lower bounds for $\|\mathscr{P}\|$ and $\|\mathscr{S}\|$

$$(18.3) \qquad 1 \le \|\underset{\sim}{w}\|_* = \|\mathscr{P}\| \ , \quad \frac{1}{\operatorname{dist}(\lambda_0, \sigma(A)\setminus\{\lambda_0\})} = r_\sigma(\mathscr{S}) \le \|\mathscr{S}\| \ ,$$

by (2.1) and (7.3). Notice that (18.1) gives a bound on the *absolute error in* λ_0. If $|\lambda_0|$ is small compared to $\|E\|$, then the relative error $|\hat{\lambda}_0 - \lambda_0|/|\lambda_0|$ can be large even if the condition number $\|\mathscr{P}\|$ is small. On the other hand (18.2) gives a bound on the *relative error* in $\underset{\sim}{u}$, since $\|\underset{\sim}{u}\| = 1$.

The relation $1/\operatorname{dist}(\lambda_0, \sigma(A)\setminus\{\lambda_0\}) \le \|\mathscr{S}\|$ shows that the unit eigenvector $\underset{\sim}{u}$ corresponding to λ_0 is ill-conditioned if λ_0 is not well-separated from the rest of the spectrum of A. Of course, in general, the condition number $\|\mathscr{S}\|$ can be large even if λ_0 is well separated from $\sigma(A) \setminus \{\lambda_0\}$.

Let us now consider the Euclidean norm on $\mathbb{C}^n$:

$$\|\underset{\sim}{x}\|_2 = (|x(1)|^2 + \ldots + |x(n)|^2)^{1/2} \ , \quad \underset{\sim}{x} \in \mathbb{C}^n \ .$$

It can be readily checked that

$$\underset{\sim}{x} = \underset{\sim}{u} - \underset{\sim}{w}/\underset{\sim}{w}^H\underset{\sim}{w}$$

is the best approximation to $\underset{\sim}{u}$ from the orthogonal complement $\{\underset{\sim}{w}\}^\perp$ of $\{\underset{\sim}{w}\}$. (See [L], 23.2.) Hence

$$(18.4) \qquad \|\mathscr{P}\|_2 = \|\underset{\sim}{w}\|_2 = \frac{1}{\|\underset{\sim}{x} - \underset{\sim}{u}\|_2} = \frac{1}{\operatorname{dist}(\underset{\sim}{u}, \{\underset{\sim}{w}\}^\perp)} \ .$$

Note that if θ denotes the acute angle between $\underset{\sim}{u}$ and $\underset{\sim}{w}$, then

$$\cos\theta = |\langle \underset{\sim}{u},\underset{\sim}{w}\rangle| / \|\underset{\sim}{u}\|_2 \|\underset{\sim}{w}\|_2 = 1/\|\mathcal{P}\|_2 .$$

Now, $\{\underset{\sim}{w}\}^{\perp}$ is the null space of the spectral projection $\mathcal{P}$ associated with A and λ_0 ; it is spanned by the generalized eigenvectors of A corresponding to its eigenvalues other than λ_0 , as we see by (7.18).

If the eigenvector $\underset{\sim}{u}$ corresponding to λ_0 is 'close' to the other generalized eigenvectors of A (i.e., $\underset{\sim}{u}$ is nearly a linear combination of them, or the acute angle θ between $\underset{\sim}{u}$ and $\underset{\sim}{w}$ is close to $\pi/2$), then the eigenvalue λ_0 is ill-conditioned.

On the other hand, the condition number for λ_0 is best when $\|\mathcal{P}\|_2 = 1$, i.e., $\mathcal{P}$ is an orthogonal projection, or $\mathcal{P}^H = \mathcal{P}$. This happens if and only if $\underset{\sim}{u}$ is orthogonal to $\{w\}^{\perp}$, i.e., $\underset{\sim}{w} = \underset{\sim}{u}$, because

$$\mathcal{P} = \underset{\sim}{u}\underset{\sim}{w}^H , \quad \underset{\sim}{w}^H\underset{\sim}{u} = 1 = \underset{\sim}{u}^H\underset{\sim}{u} .$$

In this case, it follows by Problem 8.7 that

$$\|\mathcal{P}\|_2 = \frac{1}{\sqrt{\mu}} , \tag{18.5}$$

where μ is the smallest nonzero eigenvalue of $(A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)$.

If A is a normal operator, then $\|\mathcal{P}\|_2 = 1$ (Theorem 8.4 and Proposition 2.3), $\mathcal{P}$ is normal, and by (8.14),

$$\|\mathcal{P}\|_2 = r_\sigma(\mathcal{P}) = 1/\mathrm{dist}(\lambda_0,\sigma(A)\setminus\{\lambda_0\}) . \tag{18.6}$$

Thus, for a normal operator A , the stability of the eigenvalue problem $A\underset{\sim}{u} = \lambda_0\underset{\sim}{u}$ depends solely on the distance of λ_0 from the rest of the spectrum of A .

It is interesting to note that the condition number for the *eigenvalue* λ_0 involves the 'distance' of the corresponding unit *eigenvector* $\underset{\sim}{u}$ from the remaining *generalized eigenvectors* of A, while an estimate for the condition number for the *eigenvector* $\underset{\sim}{u}$ involves the distance of the corresponding *eigenvalue* λ_0 from the remaining *eigenvalues* of A.

Eigenvector of A^H corresponding to $\bar{\lambda}_0$

Since λ_0 is a simple eigenvalue of A, $\bar{\lambda}_0$ is a simple eigenvalue of A^H (Theorem 8.2(c)). Let $\underset{\sim}{u}$ be an eigenvector of A corresponding to λ_0. We wish to find an eigenvector $\underset{\sim}{v}$ of A^H corresponding to $\bar{\lambda}_0$ such that $\underset{\sim}{v}^H\underset{\sim}{u} = 1/\lambda_0$.

In case the eigenvector $\underset{\sim}{u}$ of A corresponding to λ_0 is known to be orthogonal to all the other generalized eigenvectors of A corresponding to the remaining eigenvalues, i.e., $\underset{\sim}{u}$ is orthogonal to $\{v\}^{\perp}$, then

$$\underset{\sim}{v} = \underset{\sim}{u} / \bar{\lambda}_0\|\underset{\sim}{u}\|^2 .$$

Hence there is no need to do any further work. In the absence of the knowledge of the orthogonality of $\underset{\sim}{u}$ to all the other generalized eigenvectors of A, there are two ways of proceeding to find $\underset{\sim}{v}$.

Firstly, we can solve the eigenvalue problem for A^H. Observing that $\bar{\lambda}_0$ is one of the eigenvalues of A^H, we pick a unit eigenvector $\underset{\sim}{u}'$ of A^H corresponding to $\bar{\lambda}_0$, and let $\underset{\sim}{v} = \underset{\sim}{u}' / \bar{\lambda}_0\underset{\sim}{u}'^H\underset{\sim}{u}$. (Note: $\underset{\sim}{u}'^H\underset{\sim}{u}' \neq 0$, by (8.7).) As before, the condition number for finding $\underset{\sim}{u}'$ in this manner is $\|\mathscr{S}'\|$, where $\mathscr{S}'$ is the reduced resolvent associated with A^H and $\bar{\lambda}_0$. But $\mathscr{S}' = \mathscr{S}^*$ by (8.4), where $\mathscr{S}$ is the reduced resolvent associated with A and λ_0. Thus, the desired condition number is $\|\mathscr{S}^*\| = \|\mathscr{S}\|$, as earlier.

Alternatively, (8.7) shows that there is a unique eigenvector $\underset{\sim}{w}$ of A^H corresponding to $\bar{\lambda}_0$ which satisfies $\underset{\sim}{u}^H\underset{\sim}{w} = 1$. Then $\underset{\sim}{v}$ is the unique solution of the system

$$\begin{array}{ll} & \underset{\sim}{u}^H\underset{\sim}{x} = 1/\bar{\lambda}_0 \\ (18.7) & \\ & (A^H-\bar{\lambda}_0 I)\underset{\sim}{x} = \underset{\sim}{0} \, . \end{array}$$

of $(n+1)$ equations in the n unknowns $x(1),\ldots,x(n)$. Solving this system is much simpler than solving the eigenvalue problem for A^H .

Consider the $(n+1) \times n$ coefficient matrix

$$(18.8) \qquad \bar{C} = \begin{array}{c} \left[\begin{array}{c} \overline{u(1)} \; \ldots \; \overline{u(n)} \\ \\ A^H - \bar{\lambda}_0 I \end{array}\right] \end{array} \begin{array}{l} 1 \\ \\ n \end{array}$$

of the system (18.7). Since $\bar{\lambda}_0$ is a simple eigenvalue of A^H and $\underset{\sim}{u}$ is an eigenvector of $(A^H)^H = A$ corresponding to $\bar{\bar{\lambda}}_0 = \lambda_0$, it follows by (8.7) that $\bar{C}\underset{\sim}{x} = 0$ implies $\underset{\sim}{x} = 0$. Thus, the map $\bar{C} : \mathbb{C}^n \to \mathbb{C}^{n+1}$ is one to one, i.e., the matrix $\bar{C}$ has rank n . The system then can be solved by reducing $\bar{C}$ to an upper trapezoidal form either by the Householder orthogonalization method or by Gaussian elimination method with partial pivoting. (See Theorems 3 and 2 as well as other relevant comments in Appendix II.) The unique solution of the linear system (18.7) is given by

$$\underset{\sim}{v} = \bar{C}^{\dagger}[1/\bar{\lambda}_0, 0, \ldots, 0]^t \, ,$$

where $\bar{C}^{\dagger} : \mathbb{C}^{n+1} \to \mathbb{C}^n$ is the <u>Moore-Penrose inverse</u> of $\bar{C} : \mathbb{C}^n \to \mathbb{C}^{n+1}$:

$$\bar{C}^{\dagger} = (\bar{C}^H\bar{C})^{-1}\bar{C}^H \, .$$

(See Appendix II, especially (5).) For every $\underset{\sim}{y} \in \mathbb{C}^n$, there is a unique $\underset{\sim}{x} \in \mathbb{C}^{n+1}$ (known as the least squares solution) such that

$$\min_{\underset{\sim}{z}\in\mathbb{C}^n} \|\bar{C}\underset{\sim}{z} - \underset{\sim}{y}\|_2 = \|\bar{C}\underset{\sim}{x} - \underset{\sim}{y}\|_2 .$$

Theorem II.6 shows that the relative change in the (least squares) solution $\underset{\sim}{y}$ due to perturbations of $\bar{C}$ and $[1/\bar{\lambda}_0,0,\ldots,0]$ depends on the condition number

$$k_2(\bar{C}) = \|\bar{C}\|_2 \, \|\bar{C}^{\dagger}\|_2 .$$

We have

$$\bar{C}^H\bar{C} = \underset{\sim}{u}\underset{\sim}{u}^H + (A-\lambda_0 I)(A^H-\bar{\lambda}_0 I) . \tag{18.9}$$

Let $\sigma_1(\bar{C})$ (resp., $\sigma_n(\bar{C})$) denote the positive square root of the largest (resp., smallest) eigenvalue of $\bar{C}^H\bar{C}$. Then

$$k_2(\bar{C}) = \sigma_1(\bar{C}) \,/\, \sigma_n(\bar{C}) ,$$

by (7) of Appendix II. This perturbation analysis is applicable to the round-off errors that arise while solving the linear system (18.7) by the Householder orthogonalization method. (See (20) of Appendix II.)

Solution of the system $\underset{\sim}{v}^H\underset{\sim}{x} = 0$, $(A-\lambda_0 I)\underset{\sim}{x} = \underset{\sim}{\beta}$.

We now take up the last question regarding numerical stability that arises while implementing the algorithms of Section 17. Let $\underset{\sim}{u}$ be an eigenvector of A corresponding to a nonzero simple eigenvalue λ_0 such that $\underset{\sim}{u}^H\underset{\sim}{u} = \|u\|_2^2 = 1$. Let $\underset{\sim}{v}$ be the eigenvector of A^H corresponding to $\bar{\lambda}_0$ such that $\underset{\sim}{v}^H\underset{\sim}{u} = 1/\lambda_0$. The following linear system occurs in the calculation of the eigenvector iterates φ_j , $j = 1,2,\ldots$:

$$\begin{aligned} &\underset{\sim}{v}^H\underset{\sim}{x} = 0 \\ &(A-\lambda_0 I)\underset{\sim}{x} = \underset{\sim}{\beta} , \end{aligned} \tag{18.10}$$

where $\underset{\sim}{\beta} \in \mathbb{C}^n$ satisfies $\underset{\sim}{v}^H\underset{\sim}{\beta} = 0$. (See, e.g., Step 2(ii) of Algorithm

17.8.) While the right hand side $\underset{\sim}{\beta}$ changes from iterate to iterate, the coefficient matrix

$$(18.11) \qquad C = \begin{array}{c} \left[\begin{array}{c} \overline{v(1)} \; \dots \; \overline{v(n)} \\ \\ A - \lambda_0 I \end{array}\right] \begin{array}{l} 1 \\ \\ n \end{array} \\ n \end{array}$$

of the system remains unchanged throughout the iteration process.

The spectral projection associated with A and λ_0 is given by $\mathscr{P} = \lambda_0 \underset{\sim}{u}\underset{\sim}{v}^H$, so that

$$Z(\mathscr{P}) = \{\underset{\sim}{\beta} \in \mathbb{C}^n : \underset{\sim}{v}^H \underset{\sim}{\beta} = 0\} .$$

Since λ_0 is simple (and hence semisimple), it follows from Lemma 7.1(b) that $R(A-\lambda_0 I) = Z(\mathscr{P})$. Also, the operator $(A-\lambda_0 I)|_{Z(\mathscr{P})}$ is invertible. Thus, for every $\underset{\sim}{\beta} \in \mathbb{C}^n$ satisfying $\underset{\sim}{v}^H \underset{\sim}{\beta} = 0$, the system (18.10) has a unique solution $\underset{\sim}{x} \in \mathbb{C}^n$; in fact, $\underset{\sim}{x} = \mathscr{S}\underset{\sim}{\beta}$, where $\mathscr{S}$ is the reduced resolvent associated with A and λ_0 . Again, the solution can be calculated by reducing the system (18.10) to an upper trapezoidal form by the Householder orthogonalization method or Gaussian elimination with partial pivoting. (See Appendix II.)

Now, the matrix C has rank n . For an arbitrary $\underset{\sim}{y} \in \mathbb{C}^n$, the unique least squares solution of the system (18.10) is given by $C^\dagger[0,y(1),\dots,y(n)]^t$, where $C^\dagger = (C^H C)^{-1} C^H$ is the Moore-Penrose inverse of C . As in the case of the matrix $\bar{C}$ given by (18.8), we see by (7) of Appendix II that the condition number for the (least squares) solution $\underset{\sim}{x}$ of $C\underset{\sim}{x} = \underset{\sim}{\beta}$, where $\underset{\sim}{v}^H \underset{\sim}{\beta} = 0$, is

$$k_2(C) = \sigma_1(C) \,/\, \sigma_n(C) ,$$

where $\sigma_1(C)$ (resp., $\sigma_n(C)$) is the positive square root of the largest (resp., smallest) eigenvalue of $C^H C$. Note that

$$C^HC = \underset{\sim}{v}\underset{\sim}{v}^H + (A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)$$

$$(18.12) \qquad = \frac{\mathscr{P}^H\mathscr{P}}{|\lambda_0|^2} + (A^H-\bar{\lambda}_0 I)(A-\lambda_0 I) \ .$$

Assume now that $\mathscr{P}^H = \mathscr{P}$. (We have seen earlier that this case arises when the eigenvector $\underset{\sim}{u}$ of A corresponding to λ_0 is orthogonal to all the generalized eigenvectors of A corresponding to the remaining eigenvalues of A . This is certainly true if A is normal.) We then have

$$C^HC = \frac{\mathscr{P}}{|\lambda_0|^2} + (A^H-\bar{\lambda}_0 I)(A-\lambda_0 I) \ .$$

Since $(A-\lambda_0 I)\mathscr{P} = 0 = \mathscr{P}^H(A^H-\bar{\lambda}_0 I)$, we see that C^HC commutes with $\mathscr{P}$. If we let $Y = R(\mathscr{P})$ and $Z = Z(\mathscr{P})$, then by (6.2)

$$\sigma(C^HC) = \sigma(C^HC|_Y) \cup \sigma(C^HC|_Z) \ .$$

Now, $C^HC|_Y = \frac{1}{|\lambda_0|^2} I|_Y$ and $C^HC|_Z = (A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)|_Z$. Also, $R((A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)) = R(A^H-\bar{\lambda}_0 I) = Z(\mathscr{P}^H) = Z(\mathscr{P}) = Z$, and $Z((A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)) = Z(A-\lambda_0 I) = R(\mathscr{P}) = Y$. Hence

$$\sigma(C^HC) = \left\{\frac{1}{|\lambda_0|^2}\right\} \cup \left\{\mu : 0 \neq \mu \in \sigma((A^H-\bar{\lambda}_0 I)(A-\lambda_0 I))\right\} .$$

Let $\mu_1 \geq \mu_2 \geq \cdots \mu_{n-1}$ be the nonzero eigenvalues of $(A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)$. Then

$$\sigma_1(C) = \max\left\{\sqrt{\mu_1} \ , \ 1/|\lambda_0|\right\} , \ \sigma_n(C) = \min\left\{\sqrt{\mu_{n-1}} \ , \ 1/|\lambda_0|\right\} ,$$

$$(18.13) \qquad k_2(C) = \frac{\max\{\sqrt{\mu_1} \ , \ 1/|\lambda_0|\}}{\min\{\sqrt{\mu_{n-1}} \ , \ 1/|\lambda_0|\}} \ .$$

Now, since $(A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)$ is self-adjoint, we have

$$\mu_1 = \|(A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)\|_2 = \|A-\lambda_0 I\|_2^2 \ .$$

Also, since $\mathscr{S}$ is orthogonal, we see by Problem 8.7 that

$$\mu_{n-1} = 1/\|\mathscr{S}\|_2^2 \ .$$

Hence

$$k_2(C) = \max\left\{\|A-\lambda_0 I\|_2, \frac{1}{|\lambda_0|}\right\} \max\{\|\mathscr{S}\|_2, |\lambda_0|\} \ . \tag{18.14}$$

In particular, let A be normal and let $\lambda_1,\ldots,\lambda_{n-1}$ be the eigenvalues of A other than λ_0, arranged so that $|\lambda_1-\lambda_0| \geq \ldots \geq |\lambda_{n-1}-\lambda_0|$. Then $\bar{\lambda}_1,\ldots,\bar{\lambda}_{n-1}$ are eigenvalues of A^H, and we have $\mu_i = |\lambda_i-\lambda_0|^2$, $i = 1,\ldots,n-1$. Thus,

$$k_2(C) = \frac{\max\{|\lambda_1-\lambda_0| \ , \ 1/|\lambda_0|\}}{\min\{|\lambda_{n-1}-\lambda_0| \ , \ 1/|\lambda_0|\}} \quad (A \text{ normal}) \ . \tag{18.15}$$

Thus, in this case, the condition number of C depends on $1/|\lambda_0|$ and on the distances from λ_0 to the remaining eigenvalues of A.

Finally, we remark that the first equation $\underset{\sim}{v}^H \underset{\sim}{x} = 0$ of the system (18.10) can be scaled by multiplying it by a constant $\zeta \neq 0$, without affecting the solution of the system. In that case, the coefficient matrix

$$C_\zeta = \begin{bmatrix} \zeta\overline{v(1)} & \ldots & \zeta\overline{v(n)} \\ & & \\ & A - \lambda_0 I & \end{bmatrix} \tag{18.16}$$

of the scaled system has the condition number

$$\begin{aligned} k_2(C_\zeta) &= \frac{\max\{\sqrt{\mu_1}, |\zeta/\lambda_0|\}}{\min\{\sqrt{\mu_{n-1}}, |\zeta/\lambda_0|\}} \\ &= \max\{\|A-\lambda_0 I\|_2, |\zeta/\lambda_0|\}\max\{\|\mathscr{S}\|_2, |\lambda_0/\zeta|\} \ . \end{aligned} \tag{18.17}$$

If we choose the scaling factor such that $|\zeta/\lambda_0|$ equals (or is close to) either $\sqrt{\mu_1}$ or $\sqrt{\mu_{n-1}}$, then $k_2(C_\zeta)$ equals (or is close to) $\sqrt{\mu_1} / \sqrt{\mu_{n-1}} = \|A-\lambda_0 I\|_2\|\mathscr{S}\|_2$. Thus, $k_2(C_\zeta)$ is smallest if the first row of C is changed to

$$\lambda_0\sqrt{\mu_1}\ [\overline{v(1)},\ldots,\overline{v(n)}]\ ,$$

where $\underset{\sim}{v}$ is an eigenvector of A^H corresponding to $\bar{\lambda}_0$ such that $\lambda_0\underset{\sim}{v}^H\underset{\sim}{u} = 1 = \underset{\sim}{u}^H\underset{\sim}{u}$ ..

In case $\mathcal{P}^H \neq \mathcal{P}$, it is not clear how $k_2(C_\zeta)$ depends on $\|A-\lambda_0 I\|_2$, $\|\mathcal{S}\|_2$ and $\|\mathcal{P}\|_2$ in a precise manner. Here is a most simple-minded example. Let $A = \begin{bmatrix} 1+\lambda_0 & a \\ 0 & \lambda_0 \end{bmatrix}$, $a \in \mathbb{C}$. Let $\underset{\sim}{u} = \begin{bmatrix} -a \\ 1 \end{bmatrix} / \sqrt{1+|a|^2}$, $\underset{\sim}{v} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}\sqrt{1+|a|^2} / \bar{\lambda}_0$, so that $\mathcal{P} = \begin{bmatrix} 0 & -a \\ 0 & 1 \end{bmatrix}$ and $\mathcal{S} = \begin{bmatrix} 1 & a \\ 0 & 0 \end{bmatrix}$. Then $\|A-\lambda_0 I\|_2 = \|\mathcal{P}\|_2 = \|\mathcal{S}\|_2 = \sqrt{1+|a|^2}$. On the other hand, with $t = \zeta/\lambda_0$, we obtain

$$C_\zeta^H C_\zeta = \begin{bmatrix} 1 & a \\ \bar{a} & |a|^2+|t|^2+|a|^2|t|^2 \end{bmatrix} ,\quad \text{so that}$$

$$k_2(C_\zeta) = \frac{1+|t|^2}{2|t|}\left[(1+|a|^2)^{1/2} + \left[|a|^2 + \frac{(1-|t|^2)^2}{(1+|t|^2)^2}\right]^{1/2}\right] .$$

For $\zeta = \lambda_0$, we have

$$k_2(C_{\lambda_0}) = \sqrt{1+|a|^2} + |a| .$$

We now describe another way of finding $\underset{\sim}{x} \in \mathbb{C}^n$ such that

$$(18.10) \qquad \underset{\sim}{v}^H\underset{\sim}{x} = 0 \quad \text{and} \quad (A-\lambda_0 I)\underset{\sim}{x} = \underset{\sim}{\beta} ,$$

where $\underset{\sim}{v}^H\underset{\sim}{\beta} = 0$. Consider

$$(18.18) \qquad B \equiv -\lambda_0\mathcal{P} + (A-\lambda_0 I) = A - \lambda_0(I + \lambda_0\underset{\sim}{u}\underset{\sim}{v}^H) .$$

(Note: $\mathcal{P} = \lambda_0\underset{\sim}{u}\underset{\sim}{v}^H$.) We show that $\underset{\sim}{x}$ satisfies (18.10) if and only if $B\underset{\sim}{x} = \underset{\sim}{\beta}$: If $\underset{\sim}{x}$ satisfies (18.10), then

$$B\underset{\sim}{x} = A\underset{\sim}{x} - \lambda_0(\underset{\sim}{x} + \underset{\sim}{0}) = A\underset{\sim}{x} - \lambda_0\underset{\sim}{x} = \underset{\sim}{\beta} \ .$$

Conversely, let $B\underset{\sim}{x} = \underset{\sim}{\beta}$, i.e.,

(18.19) $$A\underset{\sim}{x} - \lambda_0\underset{\sim}{x} - \lambda_0^2\underset{\sim}{u}\underset{\sim}{v}^H\underset{\sim}{x} = \underset{\sim}{\beta} \ .$$

Since $(A^H - \bar{\lambda}_0 I)\underset{\sim}{v} = \underset{\sim}{0}$, we have $\underset{\sim}{v}^H(A-\lambda_0 I) = \underset{\sim}{0}$. Also, $\underset{\sim}{v}^H\underset{\sim}{u} = 1/\lambda_0$. Hence taking inner products with $\underset{\sim}{v}$ on both sides of (18.19), we obtain

$$-\lambda_0\underset{\sim}{v}^H\underset{\sim}{x} = \underset{\sim}{v}^H\underset{\sim}{\beta} = \underset{\sim}{0} \ .$$

Since $\lambda_0 \neq 0$, this implies $\underset{\sim}{v}^H\underset{\sim}{x} = 0$. Also, (18.19) gives

$$A\underset{\sim}{x} - \lambda_0\underset{\sim}{x} = \underset{\sim}{\beta} \ ,$$

as desired.

Now, the operator B commutes with the projection $\mathscr{P}$. With $Y = R(\mathscr{P})$ and $Z = Z(\mathscr{P})$, it follows by (6.2) that

$$\sigma(B) = \sigma(B|_Y) \cup \sigma(B|_Z) \ .$$

But $B|_Y = -\lambda_0 I|_Y$, so that $\sigma(B|_Y) = \{-\lambda_0\}$. Also, $B|_Z = (A-\lambda_0 I)|_Z$, and the spectral decomposition theorem (Theorem 6.3) gives

$$\sigma(B|_Z) = \{\lambda-\lambda_0 \ : \ \lambda \in \sigma(A) \ , \ \ \lambda \neq \lambda_0\} \ .$$

Let, as before, $\lambda_1,\ldots,\lambda_{n-1}$ be the eigenvalues of A other than λ_0 , arranged so that $|\lambda_1-\lambda_0| \geq \ldots \geq |\lambda_{n-1}-\lambda_0|$. Then

(18.20) $$\sigma(B) = \{\lambda_0\} \cup \{\lambda_1-\lambda_0,\ldots,\lambda_{n-1}-\lambda_0\} \ .$$

Since $\lambda_0 \neq 0$ and $\lambda_0 \neq \lambda_i$, $i = 1,\ldots,n-1$, we see that B is invertible. The solution $\underset{\sim}{x}$ of (18.10) can thus be obtained by solving $B\underset{\sim}{x} = \underset{\sim}{\beta}$, by Gaussian elimination with partial pivoting or by the Householder orthogonalization method, the condition number for the solution being

$$k(B) = \|B\| \, \|B^{-1}\| .$$

For the Euclidean norm $\| \, \|_2$, we have

$$k_2(B) = \frac{\sigma_1(B)}{\sigma_n(B)} ,$$

where $\sigma_1(B)$ (resp., $\sigma_n(B)$) is the largest (resp., smallest) eigenvalue of B^HB . Now,

$$B^HB = |\lambda_0|^2 \mathcal{S}^H\mathcal{S} + (A^H-\bar{\lambda}_0 I)(A-\lambda_0 I) - E ,$$

where

$$E = \lambda_0(A^H-\bar{\lambda}_0 I)\mathcal{S} + \bar{\lambda}_0 \mathcal{S}^H(A-\lambda_0 I) .$$

Recalling (18.12) and (18.16), we see that

$$B^HB = C_\zeta^H C_\zeta - E , \quad \text{where} \quad \zeta = |\lambda_0|^2 .$$

If $\mathcal{S}$ is orthogonal, then $(A^H-\bar{\lambda}_0 I)\mathcal{S} = (A^H-\bar{\lambda}_0 I)\mathcal{S}^H = 0$ and similarly $\mathcal{S}^H(A-\lambda_0 I) = \mathcal{S}(A-\lambda_0 I) = 0$, so that $E = 0$. Thus, $B^HB = C_\zeta^H C_\zeta$ with $\zeta = |\lambda_0|^2$, and the stability considerations are exactly as before. In particular,

$$(18.21) \qquad k_2(B) = \frac{\max\{\sqrt{\mu_1} \, , \, |\lambda_0|\}}{\min\{\sqrt{\mu_{n-1}}, |\lambda_0|\}} = \max\{\|A-\lambda_0 I\|_2, |\lambda_0|\}\max\{\|\mathcal{S}\|_2, 1/|\lambda_0|\},$$

where μ_1 is the largest eigenvalue of $(A^H-\bar{\lambda}_0 I)(A-\lambda_0 I)$, and μ_{n-1} is its smallest nonzero eigenvalue. (See (18.17).) In case A is normal, then $\mu_1 = |\lambda_1-\lambda_0|^2$ and $\mu_{n-1} = |\lambda_{n-1}-\lambda_0|^2$. This result also follows directly if we note that B is normal and use (18.20).