Perturbation Analysis

(Backward error)

Since we cannot compute exactly we are concerned with effects of errors. (usually relative errors)

Consider comp. $f(x)$ exact: $y = f(x)$
comp: $\tilde{y}$

In general we want to know $\|\tilde{y} - y\|$ or $\frac{\|\tilde{y}-y\|}{\|y\|}$ (forward error). It turns out that this approach has problems (soon). Instead we assume we computed exact answer to perturbed problem (input) and assess the effect of that perturbation by perturbation analysis (backward error = perturbation, it may not be unique)

computed exactly $f(x+\varepsilon)$ (or $f(x+\varepsilon) + \Delta f(x+\varepsilon)$)

backward error is ε; analyze $|f(x+\varepsilon) - f(x)|$

Example: Compute $u^T v$ $\quad u, v \in \mathbb{R}^n$
↳ u_i, v_i machine numbers

exact answer: $\sum u_i v_i$

Floating pt. $fl(u^T v) = (u_1 v_1(1+\varepsilon_1) + u_2 v_2(1+\varepsilon_2))(1+\varepsilon_3) + \cdots$

usually we just write $((u_1 v_1(1+\varepsilon) + u_2 v_2(1+\varepsilon))(1+\varepsilon) + u_3 v_3(1+\varepsilon))(1+\varepsilon) \cdots$

and keep in mind that all "ε" are different but $|\varepsilon| \le \varepsilon_M$

$$fl(u^Tv) = u_1v_1(1+\varepsilon)^n + u_2v_2(1+\varepsilon)^n + u_3v_3(1+\varepsilon)^{n-1} + \cdots$$

(if u_i, v_i are not mach. numbers it adds factor $(1+\varepsilon)^2$ to each term)

$u_1 \to u_1(1+\varepsilon)$ $u_2 \to u_2(1+\varepsilon)$ etc

$\to (\varepsilon_1+\varepsilon_2+\cdots)$

$$fl(u^Tv) \cong u^Tv + u_1v_1 n\varepsilon + u_2v_2 n\varepsilon + \cdots$$

$$|fl(u^Tv) - u^Tv| \leq n|u|^T|v|\varepsilon$$

* sign of each ε may differ $\to$ assume all errors accumulate (worst case)

* assume ε sufficiently small (relative to n) that we can ignore $\varepsilon^2, \varepsilon^3, \ldots$ terms.

* simplify by taking $n\varepsilon|u_3v_3|$

i.e. $(n-1)\varepsilon|u_3v_3|$ etc

$$\text{relative error} \quad \frac{n\varepsilon|u|^T|v|}{|u^Tv|} = \frac{|fl(u^Tv) - u^Tv|}{|u^Tv|}$$

rel. error can be huge if $|u^Tv|$ is very small compared with $|u|^T|v|$ even if $n\varepsilon$ still small.

$\|u\|_\infty = \max_i |u_i| = 1$ $\|v\|_\infty = 1$

but $u^Tv \cong 0$ (orthogonal)

So, there is no bound on (forward) relative error

We may conclude that "simple" computation of dot product is therefore unreliable.

Backward error analysis:

$$fl(u^T v) \simeq u^T v + \cdots$$

$$\simeq n\varepsilon u_1 v_1 + n\varepsilon u_2 v_2 + \cdots$$

$$= u_1 v_1 (1+n\varepsilon) + u_2 v_2 (1+n\varepsilon) + u_3 v_3 (1+n\varepsilon) + \cdots$$

$$\simeq u_1 (1+\tfrac{1}{2}n\varepsilon)\, v_1 (1+\tfrac{1}{2}n\varepsilon) + u_2 (1+\tfrac{1}{2}n\varepsilon) v_2 (1+\tfrac{1}{2}n\varepsilon) + \cdots$$

(assuming $\frac{1}{4}n^2\varepsilon^2$ negligible)

$$= \tilde{u}_1^T \tilde{v}_1 \quad \text{(exactly)}$$

$\tilde{u}_i = u_i + \eta$ where $|\eta_i| \le \frac{1}{2} n \varepsilon_M$

so $\|\eta\|_\infty \le \frac{1}{2} n \varepsilon_M$ and $\|u\|_\infty = 1$

So, floating pt. computation of dot product has small relative backward error.

What explains the huge relative forward error of $u^T v \simeq 0$ is the "sensitivity" or "conditioning" of the problem. If $u^T v \simeq 0$ the dot product is ill-conditioned. If $u^T v$ not very small the problem is well-conditioned.

Backward error analysis let's us separate the accumulation of numerical error from

the sensitivity of the problem. So, we combine backw. error anal. with perturbation analysis to obtain bounds on computed answers. If an algorithm produces small (bounded) backward error the accuracy depends on the sensitivity

ill-cond → answer may be poor (but cannot be helped)

well-cond → answer accurate.

Perturbation Theory

The eig. vals/vec.s are (typically) exact for slightly perturbed matrix (small backward error)

We want to know how much eig. val.s and vec.s change under small perturbation

1) Eigenvalues are continuous function(s) of the matrix (coefficients)

Let A have eig. values $d_1 \cdots d_n$ (counting mult.s) and $\tilde{A} = A + E$ have eig. values $\tilde{d}_1, \cdots, \tilde{d}_n$. Then there exists ordering $j_1 \cdots j_n$ s.t.

$$|d_i - \tilde{d}_{j_i}| \leq 4\left(\|A\|_2 + \|\tilde{A}\|_2\right)^{1-\frac{1}{n}} \|E\|_2^{\frac{1}{n}}$$

Theo 3.1 p. 37

$$\text{sep}(A, \tilde{A}) \leq (2n-1)\cdot \|A\|_2^{1-\frac{1}{n}} \cdot \|E\|_2^{\frac{1}{n}}$$

"Proof" (From Stewart and Sun)

$$hd(A, \tilde{A}) \leq \left(\|A\|_2 + \|\tilde{A}\|_2\right)^{1-\frac{1}{n}} \|E\|^{\frac{1}{n}}$$

$$md(A, \tilde{A}) \leq (2n-1)\cdot hd(A, \tilde{A})$$

* The exponent $\frac{1}{n}$ is in general pessimistic but necessary (sharp → obtained in specific cases, esp. large Jordan block)

Gershgorin disks

$Ax = \lambda x$, let x_i be largest abs. comp. scale x s.t. $x_i = 1$ $(|x_j| \leq 1 \quad j \neq i)$

ith row for $Ax = \lambda x$

$$a_{ii} + \sum_{j \neq i} a_{ij} x_j = \lambda x_i = \lambda \iff$$

$$a_{ii} - \lambda = -\sum_{j \neq i} a_{ij} x_j$$

$$|a_{ii} - \lambda| \leq \sum_{j \neq i} |a_{ij}||x_j| \leq \sum_{j \neq i} |a_{ij}|$$

So, λ lies inside disk $G_i = \left\{ z \,\middle|\, |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}$

Define disks $G_1 \cdots G_n$ this way

Theo 3.2

Then $\Lambda(A) \subset \bigcup_{i=1,n} G_i$

Moreover if $G_{i_1} \cup \cdots \cup G_{i_k} \cap G_{i_{k+1}} \cdots G_{i_n} = \emptyset$

↑ disjoint from complement for some permutation $i_1 \cdots i_n$ of $\{1, \ldots, n\}$, then

$G_{i_1} \cup \cdots \cup G_{i_k}$ contains exactly k eigenvals

Proof by continuity $A = \underset{\uparrow}{D} + B$

diag A

$A(t) = D + tB$

$\Lambda(A(0)) = \{a_{11}, a_{22}, \ldots, a_{nn}\}$

By continuity eigenvalues stay in disks

$G_i(t) = \left\{ z : |z - a_{ii}| \leq t \sum_{j \neq i} |a_{ij}| \right\}$

If the column sum (off-diag) is much larger than the row sum we can improve the disk corresponding to that row by a diagonal similarity transformation

$D A D^{-1}$ where d_i is small

(what limits how small we can make d_i?)

Theo 3.5

Let $A = A^H$ (Hermitian) and let

$d_1 \geq d_2 \geq \cdots \geq d_n$ then

$$\max_{\substack{\dim(W)=i \\ W}} \min_{\substack{w \in W \\ \|w\|_2 = 1}} w^H A w = d_i$$

$$\min_{W:\ \dim(W) = n+1-i} \max_{\substack{w \in W \\ \|w\|_2 = 1}} w^H A w = d_i$$

Outline of Proof from Horn & Johnson I, see also Stewart & Sun (Courant-Fischer/Fischer's, Rayleigh-Ritz)

A Hermitian $\rightarrow A = U \Lambda U^H$, $\|w\|_2 = 1$

$w^H A w = w^H U \Lambda U^H w = \xi^H \Lambda \xi$ where $\xi = U^H w$. Also assume $d_1 \geq d_2 \geq \cdots \geq d_n$

$\xi^H \Lambda \xi = \sum_i d_i |\xi_i|^2$ and we have

$d_1 = \sum_i d_1 |\xi_i|^2 \geq \sum_i d_i |\xi_i|^2 = w^H A w$

note $d_1 = u_1^H A u_1$ (so bound sharp)

$\left(\sum |\xi_i|^2 = \xi^H \xi = w^H U^H U^H w = w^H w = 1 \right)$

Same way $d_n = \sum_i d_n |\xi_i|^2 \leq \sum_i d_i |\xi_i|^2 = w^H A w$

(sharp since $u_n^H A u_n = d_n$)

Consider $\max_{\substack{\|w\|_2 = 1 \\ w \perp u_1}} w^H A w$

$\xi = U^H w = \begin{pmatrix} 0 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}$ $\quad \xi_1 = 0 \Leftarrow u_1^H w = 0$

$w^H A w = \sum_i d_i |\xi_i|^2 = \sum_{i=2}^{n} d_i |\xi_i|^2 \leq d_2 \sum_{i=2}^{n} |\xi_i|^2$

$= d_2$

Hence $\max_{\substack{\|w\|_2 = 1 \\ w \perp u_1}} w^H A w = d_2$

and so on.

Now we'll show that these bounds correspond to the max/min & min/max characterizations.

Let v be arbitrary vector and consider $(v \neq 0)$

$v \in \mathbb{C}^n, v \neq 0$

"max over space $\geq$ max over subspace"

$(\Rightarrow \zeta_3 = \zeta_4 = \cdots = 0)$

1: $\max_{\substack{\|w\|_2=1 \\ w \perp v}} w^H A w$

2: $\min_{\substack{v \in \mathbb{C}^n \\ v \neq 0}} \left(\max_{\substack{\|w\|_2=1 \\ w \perp v}} w^H A w \right)$

1: $\max_{\substack{w \perp v \\ \|w\|_2=1}} w^H A w = \max \; w^H U \Lambda U^H w =$
$\cdots$

$\max_{\substack{\zeta = U^H w \perp U^H v \\ \|\zeta\|_2 = 1}} \zeta^H \Lambda \zeta = \max_{\cdots} \sum_{i=1}^{n} d_i |\zeta_i|^2$

$\geq \max_{\substack{\zeta \perp U^H v \\ |\zeta_1|^2 + |\zeta_2|^2 = 1}} d_1 |\zeta_1|^2 + d_2 |\zeta_2|^2 \geq d_2$

$\max_{\substack{\|w\|_2=1 \\ w \perp v}} w^H A w \geq d_2$ and

$\max_{\substack{w \perp u_1 \\ \|w\|_2=1}} w^H A w = d_2$

So, $\min_{v \in \mathbb{C}^n, v \neq 0} \max_{\substack{w \perp v \\ \|w\|_2=1}} w^H A w = d_2$

"retroactively" all min/max (rather than inf/sup) okay since min/max obtained (must be for min... / max... over sphere in finite dimensions)

Note that $\dim(W) = n \Rightarrow W = \mathbb{C}^n$,

$\dim(W) = n-1 \Rightarrow W \perp v \ (\neq 0)$ for some $v \in \mathbb{C}^n$

$\dim(W) = n-2 \Rightarrow W \perp v_1, v_2 \in \mathbb{C}^n$, <u>etc</u>

Based on these ideas we can prove many other properties of eigenvalues of A and of representations of A over subspace

theo 3.6 : A Hermitian, with eig. vals

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

$U^{n \times m}$ orthonormal and

$U^H A U$ has eig. vals $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$

$$\lambda_1 \geq \mu_1 \geq \lambda_{n-m+1}$$

$$\lambda_2 \geq \mu_2 \geq \lambda_{n-m+2}$$

$$\lambda_i \geq \mu_i \geq \lambda_{n-m+i} \quad i = 1 \ldots m$$

When $m = n-1 \rightarrow$

$$\lambda_1 \geq \mu_1 \geq \lambda_2 \geq \mu_2 \geq \cdots \geq \lambda_{n-1} \geq \mu_{n-1} \geq \lambda_n$$

$\lambda_i \in \Lambda(A) : \lambda_1 \geq \lambda_2 \geq \cdots$
$\tilde{\lambda}_i \in \Lambda(\tilde{A}) : \tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots$
$\varepsilon_i \in \Lambda(E) : \varepsilon_1 \geq \varepsilon_2 \geq \cdots$

Consider A, $\tilde{A} = A + E$ Hermitian

$$x_1^H \tilde{A} x_1 = x_1^H A x_1 + x_1^H E x_1 \qquad (= \lambda_1 x_1)$$

$$\tilde{\lambda}_1 = \lambda_1 + x_1^H E x_1 \qquad \varepsilon_1 \geq x_1^H E x_1 \geq \varepsilon_n$$

$$\tilde{\lambda}_1 \leq \lambda_1 + \varepsilon_1 \quad \text{and} \quad \tilde{\lambda}_1 \geq \lambda_1 + \varepsilon_n$$

More generally (Theo 3.8)

$$\lambda_i + \varepsilon_n \le \tilde{\lambda}_i \le \lambda_i + \varepsilon_1 \qquad i = 1, \dots, n$$

So, $|\tilde{\lambda}_i - \lambda_i| \le \|E\|_2 \quad (= \varepsilon_1)$

Also (Theo 3.10 : Hoffman - Wielandt)

$$\left(\sum_i (\tilde{\lambda}_i - \lambda_i)^2\right)^{1/2} \le \|E\|_F$$

(simple eig. pairs)

$$\begin{pmatrix} y^H \\ Y^H \end{pmatrix} A (x\ X) = \begin{pmatrix} \lambda & 0 \\ 0 & M \end{pmatrix}$$

$y^H x = 1$

$Y^H X = I$

Next we consider eigenpairs :

$A \to (\lambda, x) \qquad \tilde{A} = A + E \to (\tilde{\lambda}, \tilde{x})$

$\tilde{\lambda} = \lambda + \varphi$

$\tilde{x} = x + X p \qquad$ (X complementary right eigenspace)

$$(A+E)(x + Xp) = (\lambda + \varphi)(x + Xp) \iff$$

$$Ax + Ex + AXp + EXp = \lambda x + \varphi x + \lambda Xp + \varphi Xp$$

$$AXp + Ex \simeq \lambda Xp + \varphi x$$

$\begin{cases} Ax = \lambda x \\ \text{neglect "small" } \times \text{ "small"} \end{cases}$

I: $y^H AXp + y^H Ex = \lambda y^H Xp + \varphi y^H x$

$$y^H Ex = \varphi$$

II $y^H AXp + y^H Ex = \lambda y^H Xp + \varphi y^H x$

$$Mp + y^H Ex = \lambda p \iff (M - \lambda I) p = y^H Ex \iff$$

$$p = (M - \lambda I)^{-1} y^H Ex$$

$$(\tilde{\lambda}, \tilde{x}) \simeq (\lambda + y^H Ex,\ x + X(M - \lambda I)^{-1} y^H Ex)$$

Theo 3.11

Let (d, x) be simple eigen pair of A, (x, X) nonsingular and $(y\ Y)^H = (x\ X)^{-1}$

$\tilde{A} = A + E$. Let

$$\begin{pmatrix} y^H \\ Y^H \end{pmatrix} E \begin{pmatrix} x & X \end{pmatrix} = \begin{pmatrix} \varphi_{11} & f_{12}^H \\ f_{21} & F_{22} \end{pmatrix}$$

$\|\cdot\|$ is a consistent norm and

$$sep(d, M) = \|(dI - M)^{-1}\|^{-1}$$

If $4\|f_{21}\|\|f_{12}^H\| < \left(sep(d, M) - |\varphi_{11}| - \|F_{22}\|\right)^2$

then there exists φ and p such that

$(\tilde{d}, \tilde{x}) = (d + \varphi, x + Xp)$ is an eig. pair of $\tilde{A}$ with $\tilde{\lambda} \notin \Lambda(M)$

Moreover,

$$\|p\| < \frac{2\|f_{21}\|}{sep(d, M) - |\varphi_{11}| - \|F_{22}\|}$$

$$\|p - (dI - M)^{-1} f_{21}\| < \frac{2\|p\|^2 \|f_{12}^H\|}{sep(d, M) - |\varphi_{11}| - \|F_{22}\|}$$

and

$$\|\varphi - y^H E x\| \leq \|p\|\,\|f_{12}\|$$

Proof see Stewart & Sun (discuss dep. on time)

Note:

$$|\varphi| = |\tilde{d} - d| = O(\|E\|) \text{ and}$$

$$\|\varphi - y^H E x\| = O(\|E\|^2)$$

So, simple eigenvalue is much better behaved than ~~a~~ defect. eig. val.

$O(\|E\|)$ vs $O(\|E\|^{1/k})$

Moreover

$\tilde{\lambda} - y^H \tilde{A} x = \tilde{\lambda} - \lambda - y^H E x$

So, $|\tilde{\lambda} - y^H \tilde{A} x| = O(\|E\|^2)$

$y^H \tilde{A} x / y^H x$ Rayleigh quotient

↳ $= y^H A x$ (since $y^H x = 1$)

let $E = \varepsilon \ell_i \ell_j^H$ then

$\tilde{\lambda} - \lambda = \varepsilon \bar{y}_i x_j + O(\varepsilon^2)$

Hence $\lambda(a_{ij})$ differentiable and

$\frac{\partial \lambda}{\partial a_{ij}} = \bar{y}_i x_j$

eigenvector $\tilde{x}$: $\|\tilde{x} - x\| = O(\|E\|)$ and

$\|\tilde{x} - [x + X(\lambda I - M)^{-1} f_{2i}]\| = O(\|E\|^2)$

Conditioning/ Sensitivity (for small pert.)

$\tilde{\lambda} = \lambda + y^H E x + (\varphi - y^H E x)$

$|\tilde{\lambda} - \lambda| \leq \|x\| \|y\| \|E\| + O(\|E\|^2)$

$\leq \sec \Delta(x,y) \|E\| + O(\|E\|^2)$

We call $\sec \Delta(x,y)$ condition number

Note $\cos \Delta(x,y) = \left| \left(\frac{x}{\|x\|}\right)^H \left(\frac{y}{\|y\|}\right) \right| =$

$\frac{1}{\|x\| \cdot \|y\|} \Rightarrow \|x\| \cdot \|y\| = \sec \Delta(x,y)$

Also common to take $\|x\|=1$, $\|y\|=1$ and

$$\text{cond}(d) = \frac{1}{|y^H x|}$$

X, Y not unique $\rightarrow$ assume S s.t. YS^{-H} orthonormal

replace X by XS, Π by $S^{-1}MS$

Now assume Y orthonormal.

(Lemma 3.12) Then $\sin\Delta(x,\tilde{x}) = \dfrac{\|Y^H\tilde{x}\|_2}{\|\tilde{x}\|_2}$

let $\|x\|_2 = \|\tilde{x}\|_2 = 1 \rightarrow (x\ Y)$ unitary

$$\left\| \begin{pmatrix} x^H \\ Y^H \end{pmatrix} \tilde{x} \right\|_2^2 = |x^H\tilde{x}|^2 + \|Y^H\tilde{x}\|_2^2 = 1$$

$$|x^H\tilde{x}|^2 = \cos^2\Delta(x,\tilde{x}) \Rightarrow$$

$$\|Y^H\tilde{x}\|_2^2 = \sin^2\Delta(x,\tilde{x})$$

Assume Theo 3.11 wrt $\|\cdot\|_2$: Theo 3.13

$$\sin(x,\tilde{x}) \leq \frac{\|E\|_2}{\text{sep}(\tilde{d}, M)}$$

Proof:

$$(A+E)\tilde{x} = \tilde{\lambda}\tilde{x}$$

$$Y^H A\tilde{x} - \lambda Y^H\tilde{x} = (M - \tilde{\lambda}I)Y^H\tilde{x} = -Y^H E\tilde{x}$$

~~$Y^H\tilde{x} = (M-\tilde{\lambda}I)^{-1}$~~

$$Y^H\tilde{x} = (\tilde{\lambda}I - M)^{-1} Y^H E\tilde{x}$$

$$\frac{\|Y^H\tilde{x}\|_2}{\|\tilde{x}\|_2} = \|(\tilde{\lambda}I - M)^{-1}\|_2 \, \|Y^H E\tilde{x}\|_2 / \|\tilde{x}\|_2$$

$$\frac{\|y^H \tilde{x}\|_2}{\|\tilde{x}\|_2} \leq \frac{\|E\|_2 \, \|\tilde{x}\|_2}{\|(\tilde{\lambda} I - M)^{-1}\|^{-1} \, \|\tilde{x}\|_2} = \frac{\|E\|_2}{sep(\tilde{\lambda}, M)}$$

$\uparrow$

$\sin(x, \tilde{x})$

For ~~small~~ $E \to 0$, $\tilde{\lambda} \to \lambda$. So,

$$\sin(x, \tilde{x}) \lesssim \frac{\|E\|_2}{sep(\lambda, M)}$$

So, $sep^{-1}(\lambda, M)$ is a condition number for eig. vector x

Properties of sep

sep ?

$sep(\lambda, M)$ is lower bound on separation of λ from spectrum of M.

Theo 3.14

In any consistent norm,

$$sep(\lambda, M) \leq \min_{\mu \neq \lambda \in \Lambda(M)} |\lambda - \mu|$$

$$sep(\lambda, M)^{-1} = \|(\lambda I - M)^{-1}\| \geq \rho\left((\lambda I - M)^{-1}\right) =$$

$$\max_{\mu \in \Lambda(M)} |\lambda - \mu|^{-1} \Rightarrow$$

$$sep(\lambda, M) \leq \min_{\mu \in \Lambda(M)} |\lambda - \mu|$$

However, $sep(\lambda, M)$ can be much smaller than $\min |\lambda - \mu|$ if an eigenvalue of M is ill-conditioned

Hermitian case

Since eigenvalues of Hermitian matrix are perfectly well-conditioned (cond. nr = 1)

we have

Theo 3.16

Let λ be real, M Hermitian. Then in 2-norm

$$sep(\lambda, M) = \min_{\mu \in \Lambda(M)} |\lambda - \mu|$$

If A Hermitian we can take $Y = X$ (See above), so that M Herm.

$$\sin \Delta(x, \tilde{x}) \leq \frac{\|E\|_2}{\min_{\mu \in \Lambda(M)} |\lambda - \mu|}$$

So, it's the separation of eigenvalues that counts.

Canonical angles p. 250

Let X and Y be subspaces of same dim, p. X orthonormal basis for X, Y for Y.

Then $\vartheta_i = \arccos \gamma_i$ are the canonical angles between X and Y, where γ_i are singular values of $Y^H X$.

We write $\vartheta_i(X, Y)$ for i-th can. angle in descending order.

(also $\Theta(X,Y)$) $\Theta(X, Y) = \operatorname{diag}(\vartheta_1, \dots, \vartheta_p)$

If $Y_\perp$ orthonormal basis for $Y^\perp$

(Theo 2.2) then $\sigma_i(Y_\perp^H X) = \sin(\vartheta_i)$

Also

$$\begin{pmatrix} Y^H \\ Y_\perp^H \end{pmatrix} X = \begin{pmatrix} C \\ S \end{pmatrix} \quad \text{where} \quad C^H C + S^H S = I$$

$$\sin \Delta(x, Y) = \min_{y \in Y} \|x - y\|_2$$

$$\rightarrow \|x - Y Y^H x\|_2$$

Theo 2.4

Let X and Y be orthonormal, $X^H Y = 0$,

$$Z = X + Y Q$$

$\sigma_i(Z)$ nonzero singular vals descending

$\sigma_i(Q)$ " " " descending

$\vartheta_i(X, Z)$ nonzero canon. angles between $R(X)$ and $R(Z)$ in descending order

Note $K = U\Sigma V^H$

$K^H K = V \Sigma^H \Sigma V^H$

$\Sigma^H \Sigma$: eig. val.s

V: eig. vec.s

Then $\sigma_i = \sec \vartheta_i$

$\zeta_i = \tan \vartheta_i$

$$Z^H Z = (X^H + Q^H Y^H)(X + YQ)$$

$$= X^H X + \underbrace{X^H Y Q}_{=0} + \underbrace{Q^H Y^H X}_{=0} + Q^H Q$$

$$= I + Q^H Q$$

$$\sigma_i^2 = 1 + \zeta_i^2$$

To find canon. angles $\vartheta_i(X, Z)$ we need orthonormal basis for $R(Z)$

Take $\hat{Z} = Z(Z^H Z)^{-1/2} \rightarrow \hat{Z}^H \hat{Z} = I$.

$$X^H \hat{Z} = X^H (X + YQ)(I + Q^H Q)^{-1/2}$$

$$= (I + Q^H Q)^{-1/2}$$

$\cos \vartheta_i = \frac{1}{\sigma_i} \Rightarrow \sigma_i = \frac{1}{\cos \vartheta_i} = \sec \vartheta_i$

$$\sigma_i^2 = \frac{1}{\cos^2 \vartheta_i} = \frac{\cos^2 \vartheta_i + \sin^2 \vartheta_i}{\cos^2 \vartheta_i}$$

$$= 1 + \tan^2 \vartheta_i = 1 + \zeta_i^2$$

Corollary 2.5

Let X be orthonorm. basis for simple eig. space $\mathcal{X}$ of A and Y be basis for corresponding left eig. space s.t.

$$Y^H X = I$$

Then $\sigma_i(Y) = \sec \vartheta_i(X, Y)$

and $\|Y\|_2 = \sigma_{max}(Y) = \sec \vartheta_1(X, Y)$

Residual Analysis

Approx. eigenbasis X (basis for appr. eig. space):

$$R = AX - XL$$

Best L?

Theorem 2.6

$(X \; X_\perp)$ unitary $R = AX - XL$

Then $\|R\|$ minimized for any unit. invar. norm if

$$L = X^H A X \rightarrow \text{Rayleigh quotient}$$

More generally X^I left inverse of X

then $X^I A X$ is Rayleigh quot. of A

Backward Error

Theo 2.8

X orthonormal and $R = AX - XL$

Then $E = -RX^H \rightarrow (A+E)X = XL$

$$\|E\|_2 = \|R\|_2, \quad \|E\|_F = \|R\|_F$$

If A Hermitian and $L = X^H A X$

then $E = -(RX^H + X^H R)$

$$(A+E)X = XL$$

$$\|E\|_2 = \|R\|_2, \quad \|E\|_F = \sqrt{2}\,\|R\|_F$$

Proof: by construction + taking norms

(from previous typically take $L = X^H A X$

$R = AX - XL$, $X^H X = I$

We can proceed in 2 ways

II $(A+E)X = XL$ and consider perturbation eigen pair of perturbed matrix

I Consider block deflation / similarity transform and changes to X and L

I $(X\ X_\perp)$ unitary

$$\begin{pmatrix} X^H \\ X_\perp^H \end{pmatrix} A \begin{pmatrix} X & X_\perp \end{pmatrix} = \begin{pmatrix} L & H \\ G & M \end{pmatrix}$$

$$A(X\ X_\perp) = \begin{pmatrix} XL + X_\perp G & X_\perp M + XH \end{pmatrix}$$

$$AX - XL = X_\perp G = R$$

unitarily invariant norm

$\Rightarrow \|G\| = \|R\|$ for u.i.n.

$$X_\perp^H (AX - XL) = G = X_\perp^H R$$

Sim. Transform:

$$\begin{pmatrix} I & 0 \\ -P & I \end{pmatrix} \begin{pmatrix} L & H \\ G & M \end{pmatrix} \begin{pmatrix} I & 0 \\ P & I \end{pmatrix} =$$

$$\begin{pmatrix} L+HP & H \\ G+MP-PL-PHP & M-PH \end{pmatrix} = \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}$$

P s.t. $PL - MP = G - PHP$

Define

$$sep(L, M) = \min_{\|Q\|=1} \|QL - MQ\|$$

If $4\|G\|\|H\| < sep^2(L,M)$ then

unique P exists s.t. $PL - MP = G - PHP$

and

$$\|P\| < \frac{2\|G\|}{sep(L,M)}$$

Furthermore

$$\Lambda(L + HP) \cap \Lambda(M - PH) = \emptyset$$

(details in Steward & Sun)

Proof hinges on convergence of fixpoint iteration (solving for P)

note earlier alg. for Sylvester eq.

$G_0 = G \rightarrow$ Solve $P_0 : P_0 L - M P_0 = G_0$

normalise $\|A\| \leq 1$ (by scaling)

if $\|G\|$ suff. small then $\|P_0\|$ small

then $\|H\| \leq 1$

and $\|P_0\|^2$ very small $\rightarrow$

$G_1 = G_0 - P_0 H P_0$ Solve $P_1 L - M P_1 = G_1$

Overall Sun. Transform for A:

$$\begin{pmatrix} I & 0 \\ -P & I \end{pmatrix} (X \; X_\perp)^H A (X \; X_\perp) \begin{pmatrix} I & 0 \\ P & I \end{pmatrix} = \begin{pmatrix} * & * \\ 0 & * \end{pmatrix}$$

$A\hat{X} = \hat{X}\hat{L}$ where

$$\hat{X} = X + X_\perp P \;, \quad \hat{L} = L + HP$$

From theo 2.4 (chap 4) : $\tan \vartheta_i(X, \hat{X}) = \sigma_i(P)$

$$\rightarrow \tan \vartheta_i = \sigma_i(P) < \frac{2\|G\|_F}{sep(L,M)}$$

Theo 2.12

$R = AX - XL$

$S = X^H A - L X^H$

If $4\|R\|\,\|S\| \leq sep^2(L, M)$

$(\hat{L}, \hat{X})$ exists (for A)

$A\hat{X} = \hat{X}\hat{L}$

$$\|L - \hat{L}\| < \frac{2\|R\|\,\|S\|}{sep(L, M)}$$

$$\|\tan\Theta(X, \hat{X})\| < \frac{2\|R\|}{sep(L, M)}$$

---

sep is complicated but has nice properties. Note for $\|\cdot\|_2$ sep is smallest singular value of Sylvester operator (which can be written as a matrix)

Theo 2.11 lists many properties.

Here we need that sep itself is well-conditioned

$$|sep(L+E, M+F) - sep(L, M)| \leq \|E\| + \|F\|$$

(sep defined for consistent norm)

$X^H X = I$ II

$\tilde{A}X_1 = X_1 L + R \rightarrow \tilde{A}X_1 - X_1 L = R$

$(\tilde{A} - RX_1^H)X_1 = X_1 L + R - R = X_1 L$ (exact)

$E = RX^H \rightarrow \tilde{A} = A + E$ and $AX_1 = X_1 L$

Further let $A = X_1 L Y_1^H + X_2 M Y_2^H$

$\begin{pmatrix} Y_1^H \\ Y_2^H \end{pmatrix}(A+E)(X_1\ X_2) = \begin{pmatrix} L+F_{11} & F_{12} \\ F_{21} & M+F_{22} \end{pmatrix}$

(For consistent family of norms)

$\delta = sep(L, M) - \|F_{11}\| - \|F_{22}\|$

$\leftarrow$ $sep(L+F_{11}, M+F_{22}) \leqslant \delta$

If $4\|F_{21}\|\|F_{12}\| < \delta^2$ then

P exists s.t. $\|P\| < \dfrac{2\|F_{21}\|}{\delta}$

$(\tilde{L}, \tilde{X}_1) = (L + F_{11} + F_{12}P,\ X_1 + X_2 P)$

$(\tilde{M}, \tilde{Y}_2) = (M + F_{22} - PF_{12},\ Y_2 - Y_1 P^H)$

are simple, complementary, right and left eigenpairs of $\tilde{A}$