Perturbation Analysis

(Backward error)

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

Consider comp \( P(x) \) 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 an 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 \( \hat{P}(x+\epsilon) \) (or \( \hat{P}(x+\epsilon) + aP(x+\epsilon) \))

backward error (\( \epsilon \)) to analyze \( |\hat{P}(x+\epsilon) - P(x)| \)

Example: Compute \( u^Tv \), \( u, v \in \mathbb{R}^n \)

Exact answer: \( \sum u_i^tv_i \)

Machine numbers

Floating pt. \( \hat{P}(u^Tv) = (u_1v_1(1+\epsilon_1) + u_2v_2(1+\epsilon_2))(1+\epsilon_3) + \cdots \)

Usually we just write \( (u_1v_1(1+\epsilon_1) + u_2v_2(1+\epsilon_2))(1+\epsilon) + u_3v_3(1+\epsilon_3)(1+\epsilon) \)

and keep in mind that all \( \epsilon_i \) are different

but \( |\epsilon| \leq \epsilon_h \)
\[ P(uTv) = u_1v_1(1+\varepsilon)^n + u_2v_2(1+\varepsilon)^n + u_3v_3(1+\varepsilon)^{n-1} + \ldots \]

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

\[ u_i \rightarrow u_i(1+\varepsilon), \quad u_2 \rightarrow u_2(1+\varepsilon), \quad \text{etc.} \]

\[ (c_1 + c_2 + \ldots) \]

\[ P(uTv) \approx uTv + u_1v_1\varepsilon + u_2v_2\varepsilon + \ldots \]

\[ \|PuTv\| < n\|uTv\|\varepsilon. \]

* Sign of each \( \varepsilon \) may differ → assume all errors accumulate (worst case)

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

* Simplify by forgetting taking \( n\varepsilon \)!

\[ \|uTv\| \approx \left( n-1 \right) \varepsilon \|u_3v_3\| + \ldots \]

\[ -uTv \]

Relative error:
\[ \frac{\|PuTv\|}{\|uTv\|} = \frac{|P(uTv)|}{|uTv|} \]

Rel. error can be huge if \(|uTv|\) very small compared with \(|uTv|\) even if \( n\varepsilon \) still small.

\[ \|u\|_\infty = \max_i |u_i| = 1, \quad \|v\|_\infty = 1 \]

But \( uTv \approx 0 \) (orthogonal)

So, there is no bound on (forward) error.
We may conclude that "simple" computation of dot product is therefore unreliable.

**Backward error analysis**:

\[ \Phi(\mathbf{uT v}) \approx \mathbf{uT v} + \mathbf{uT e} + \mathbf{uT e} \mathbf{e} + \ldots \]

\[ = \mathbf{u}_1 \mathbf{v}_1 (1 + \eta_1^2) + \mathbf{u}_2 \mathbf{v}_2 (1 + \eta_2^2) + \mathbf{u}_3 \mathbf{v}_3 (1 + \eta_3^2) + \ldots \]

\[ = \mathbf{u}_1 (1 + \frac{1}{2} \eta_1^2) \mathbf{v}_1 (1 + \frac{1}{2} \eta_2^2) + \mathbf{u}_2 (1 + \frac{1}{2} \eta_2^2) \mathbf{v}_2 (1 + \frac{1}{2} \eta_3^2) + \ldots \]

(assuming \( \frac{1}{2} \eta^2 \) negligible)

\[ = \mathbf{u}_1 \mathbf{v}_1 \text{ (exactly)} \]

\[ \mathbf{u}_i = \mathbf{u}_i + \eta \text{ where } |\eta_i| \leq \frac{1}{2} \eta_{\text{max}} \]

\[ \Rightarrow \| \eta \|_\infty \leq \frac{1}{2} \eta_{\text{max}} \text{ and } \| \mathbf{u} \|_\infty = 1 \]

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

What explains the huge relative forward error if \( \mathbf{u}^\top \mathbf{v} \approx 0 \) is the "sensitivity" or "conditioning" of the convex problem. If \( \mathbf{u}^\top \mathbf{v} \approx 0 \) the dot product is ill-conditioned. If \( \mathbf{u}^\top \mathbf{v} \) not very small the problem is well-conditioned.

**Backward error analysis** let's use separate the accumulation of numerical error from
the sensitivity of the problem. So, we combine backward error analysis 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 $\rightarrow$ answer may be poor (but cannot be helped)

well-cond $\rightarrow$ answer accurate.
The eigenvalues and vectors are (typically) exact for slightly perturbed matrices (small backward error).

We want to know how much eigenvalues and vectors change under small perturbation.

1) Eigenvalues are continuous functions of the matrix (coefficients).

Let $A$ have eigenvalues $d_1, \ldots, d_n$ (counting multiplicity) and $\tilde{A} = A + E$ have eigenvalues $\tilde{d}_1, \ldots, \tilde{d}_n$. Then there exists ordering $\tilde{d}_1, \ldots, \tilde{d}_n$ such that:

$$|d_i - \tilde{d}_i| \leq 4(\|A\| + \|\tilde{A}\|) \frac{\|E\|}{\|A\|}$$

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

**Gershgorin disks**

$$Ax = \lambda x \Rightarrow \text{let } x_i \text{ be largest abs comp scale } x \text{ st. } x_i = 1 \text{ (} |x_i| \leq 1 \text{ if } i \neq i)$$

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

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

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

So, $\lambda$ lies inside disk $G_i = \{z \mid |z - \lambda| \leq \sum_{j \neq i} |a_{ij}|\}$

Define disks $G_1, \ldots, G_n$ in this way:

Then $\Lambda(A) \subset U \bigcup G_i$ for $i = 1, \ldots, n$. 

Theo 3.2
Moreover if \( G_{1} \cup \cdots \cup G_{k} \) and \( A_{1} \cup \cdots \cup A_{k} \) are disjoint from complement for some \( 1 \leq i, j \leq n \), then

\[ G_{i} \cup \cdots \cup G_{k} \text{ contains exactly } k \text{ eigenvalues.} \]

**Proof by continuity** \( \Lambda = \text{diag} A \)

\[ A(t) = D + tB \]

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

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 one can improve the disk corresponding to that row by a diagonal similarity transformation

\[ DAD^{-1} \text{ where } d_i \text{ is small} \]

(what limits how small we can make \( d_i \)?)

let \( A = A^H \) (Hermitian) and let

\( d_1 \geq d_2 \geq \ldots \geq d_n \)

then

\[ \max \min \begin{array}{c} w^H A w = d_i \\ \dim(W) = i \quad w \in W \\ \|w\|_2 = 1 \end{array} \]

\[ \min \max \begin{array}{c} w^H A w = d_i \\ W: \quad w \in W \\ \dim(W) = n - i \quad \|w\|_2 = 1 \end{array} \]
A Hermitian $\to A = U \Lambda U^H$, $\|w\|_2 = 1$

$w^H A w = w^H U \Lambda U^H w = \lambda_i^H \lambda_i$ where

$\lambda_i = U^H w$. Also assume $d_2 \geq d_3 \geq \ldots \geq d_n$

$\sum_{i=1}^{n} d_i/\lambda_i \geq \sum_{i=1}^{n} d_i/\lambda_i = w^H A w$

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

$(\sum_{i=1}^{n} d_i/\lambda_i \geq \sum_{i=1}^{n} d_i/\lambda_i = w^H A w$

Same way $d_n \leq \sum_{i=1}^{n} d_i/\lambda_i = \sum_{i=1}^{n} d_i/\lambda_i = w^H A w$

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

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

$S = U^H w = \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix}$

$S_i = 0 \iff u_i^H w = 0$

$w^H A w = \sum_{i=1}^{n} d_i |s_i|^2 = \sum_{i=2}^{n} d_i |s_i|^2 \leq d_2 \sum_{i=2}^{n} |s_i|^2$

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

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 $(v \neq 0)$
\[ v \in \mathbb{C}^n, \, v \neq 0 \]

\[ 1: \max_{w} w^H A w \quad \text{subject to} \quad \|w\|_2 = 1, \quad w \perp v \]

\[ 2: \min_{v \in \mathbb{C}^n} \left( \max_{w} w^H A w \right) \quad \text{subject to} \quad \|w\|_2 = 1, \quad w \perp v \]

\[ \max_{v \in \mathbb{C}^n} w^H A w = \max_{v \perp U} w^H U^H U w = \max_{v \perp U} w^H w \quad \text{subject to} \quad \|w\|_2 = 1 \]

\[ \max_{v \in \mathbb{C}^n} w^H A w = \max_{v \perp U} w^H w = \max_{\|s\|_2 = 1} \sum_{i=1}^n d_i |s_i|^2 \]

\[ \Rightarrow \max_{v \perp U} w^H w \geq \max_{\|s\|_2 = 1} \sum_{i=1}^n d_i |s_i|^2 = d_2 \]

\[ \|s\|_2 = 1 \Rightarrow |s_1|^2 + |s_2|^2 = 1 \]

\[ \max_{v \perp U} w^H w = d_2 \quad \text{and} \quad \|w\|_2 = 1 \]

\[ \max_{v \perp U} w^H A w = d_2 \quad \text{and} \quad \|w\|_2 = 1 \]

So, \[ \min_{v \in \mathbb{C}^n} \max_{w} w^H A w = d_2 \quad \text{subject to} \quad \|w\|_2 = 1, \quad w \perp v \]

"retroactively" all \( \min \) / \( \max \) (rather than \( \text{inf} / \text{sup} \)) okay since \( \min \) / \( \max \) obtained (must be for \( \min \) / \( \max \) over sphere in finite dimensions)
Note that \( \dim(W) = n \Rightarrow W = \mathbb{C}^n, \)

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

\( \dim(W) = n-2 \Rightarrow W \perp v_1, v_2 \in \mathbb{C}^n, \) etc.

Based on these ideas we can prove many other properties of eigenvalues of \( A \) and of the representation of \( A \) over a subspace.

**Theorem 3.6**: \( A \) Hermitian, with eigenvalues \( d_1 > d_2 > \ldots > d_n \)

\( U^{\text{**}} \) orthonormal and

\( U^\text{**}AU \) has eigenvalues \( \mu_1 \geq \mu_2 \geq \ldots \geq \mu_m \)

\( d_1 \geq \mu_1 \geq d_{n-m+1} \)

\( d_2 \geq \mu_2 \geq d_{n-m+2} \)

\( d_i \geq \mu_i \geq d_{n-m+i} \quad i = 1, \ldots, m \)

When \( m = n-1 \rightarrow \)

\( d_1 \geq \mu_1 \geq d_2 \geq \mu_2 \geq \ldots \geq d_{n-1} \geq \mu_{n-1} \geq d_n \)

\( \delta_i \in \Delta(A); \lambda_i \geq \delta_1 \geq \delta_2 \geq \ldots \)

\( \delta_i \in \Delta(\tilde{A}); \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \)

\( \varepsilon_i \in \Delta(E); \varepsilon_1 \geq \varepsilon_2 \geq \ldots \)

Consider \( \tilde{A} = A + E \) Hermitian

\[ x_i^\text{**} \tilde{A} x_i = x_i^\text{**} A x_i + x_i^\text{**} E x_i \quad ( \varepsilon A x_i = \lambda_i x_i ) \]

\( \hat{\lambda}_i = \lambda_i + x_i^\text{**} E x_i \quad \varepsilon_i \geq x_i^\text{**} E x_i \geq \varepsilon_n \)

\( \hat{\lambda}_1 \leq \lambda_1 + \varepsilon_1 \quad \text{and} \quad \hat{\lambda}_1 \geq \lambda_1 + \varepsilon_n \)
More generally (Theo 3.8)
\[ d_i + \varepsilon_n \leq \tilde{d}_i \leq d_i + \varepsilon_i \quad i = 1, \ldots, n \]
So,
\[ |\tilde{d}_i - d_i| \leq \|E\|_2 \quad (= \varepsilon_i) \]

Also (Theo 3.10: Hoffman–Wielandt)
\[ \left( \sum_i (\tilde{d}_i - d_i)^2 \right)^{\frac{1}{2}} \leq \|E\|_F \]

Next we consider eigenpairs:
\[ A \rightarrow (d, x) \quad \tilde{A} = A + E \rightarrow (\tilde{d}, \tilde{x}) \]
\[ \tilde{d} = d + \varphi \]
\[ \tilde{x} = x + x_p \quad (X \text{ complementary right eigenspace}) \]

\[ (A+E)(x+x_p) = (d+\varphi)(x+x_p) \quad \Leftrightarrow \quad Ax + Ex + AXp + EXp = dx + \varphi x + dXp + \varphi Xp \]

\[ AXp + Ex \approx \lambda Xp + \varphi Xp \]

I: \[ y^HAXp + y^HEx = dy^Hxp + \varphi y^Hx \]
\[ y^HEx = \varphi \]

II \[ y^HAXp + y^HEx = dy^Hxp + \varphi y^Hx \]
\[ Mp + y^HEx = dp \quad \Leftrightarrow \quad (M-dI)p = y^HEx \Leftrightarrow \]
\[ p = (M-dI)^{-1}y^HEx \]

\[ (\tilde{d}, \tilde{x}) \approx (d+y^HEx, x + x(N-dI)^{-1}y^HEx) \]
Theo 3.11

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

\[ A = A + E. \] 

\[
\begin{pmatrix}
q_H \\
F_H
\end{pmatrix}
E(x X) = 
\begin{pmatrix}
q_{11} & q_{12} \\
q_{21} & F_{22}
\end{pmatrix}
\]

\[ \| \cdot \| \] is a consistent norm and

\[ \text{sep}(d, M) = \| (dI - M)^{-1} \|^{-1} \]

\[ \| q \|_{21} \| q_{12} \| < (\text{sep}(d, M) - \| q_{11} \| - \| F_{22} \|)^2 \]

Then there exists \( q \) and \( p \) such that

\[
(\tilde{\lambda}, \tilde{x}) = (d + q, x + xp) \]

is an eig. pair of \( A \) with \( \tilde{\lambda} \in \Lambda(M) \)

Moreover,

\[ \| p \| < \frac{2\| F_{21} \|}{\text{sep}(d, M) - \| q_{11} \| - \| F_{22} \|} \]

\[ \| p - (dI - M)^{-1} F_{21} \| < \frac{2\| p \|_2^2 \| F_{12} \|}{\text{sep}(d, M) - \| q_{11} \| - \| F_{22} \|} \]

and

\[ \| \psi - \psi^HEx \| \leq \| p \| \| F_{22} \| \]

**Note:**

\[ |q| = |d - d| = O(\| E \|) \] and

\[ \| \psi - \psi^HE2x \| = O(\| E \|^2) \]
Moreover

So, simple eigenvalue much better behaved than singular defect.

eig. val.

\[ O(\|E\|) \quad vs \quad O(\|E\|^{\frac{1}{n}}) \]

\[
\tilde{d} - y^H \tilde{A} x = \tilde{d} - d - y^H E x
\]

So, \[ |\tilde{d} - y^H \tilde{A} x| = O(\|E\|^2) \]

\[
y^H \tilde{A} x / y^H x \quad \text{Rayleigh quotient}
\]

\[
\lambda \Rightarrow y^H A x \quad (\text{since } y^H x = 1)
\]

let \( E = \tilde{e} e_i e_j^T \) then

\[
\tilde{d} - d = \tilde{e} y_j y_i + O(\tilde{\varepsilon}^2)
\]

Hence \( d(a_{ij}) \) differentiable and

\[
\frac{\partial d}{\partial a_{ij}} = y_i y_j
\]

eigenvector \( \tilde{x} \):

\[
|\tilde{x} - x| = O(\|E\|) \quad \text{and}
\]

\[
|\tilde{x} - [x + X(dI - M)^{-1} f_n]| = O(\|E\|^2)
\]

\[
\tilde{d} = d + y^H E x + (y - y^H E x)
\]

\[
|\tilde{d} - d| \leq \|x\| \|y\| \|E\| + O(\|E\|^2)
\]

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

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

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

\[
\frac{1}{\|x\| \cdot \|y\|} = \|x\| \cdot \|y\| = \sec \Delta(x,y)
\]
Also common to take \( \|x\| = 1, \|y\| = 1 \) and \( \text{cond}(d) = \frac{1}{\|y^Hx\|} \).

\( X, Y \) not unique \( \Rightarrow \) assume \( S \).

\( 2.1. y^H S^{-H} \) orthonormal

replace \( X \) by \( X S \), \( \Pi \) by \( S^{-H} M S \).

Now assume \( Y \) orthonormal.

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

let \( \|x\|_2 = \|\tilde{x}\|_2 = 1 \Rightarrow (x, y) \) unitary

\[ \left\| \left( y^H \right)^2 \tilde{x} \right\|_2^2 = |x^H \tilde{x}|^2 + \|Y^H \tilde{x}\|_2^2 = 1 \]

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

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

Assume Theo 3.11; Theo 3.13

\( \text{Proof:} \)

\( (A + E) \tilde{x} = \tilde{I} \tilde{x} \)

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

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

\[ \frac{\|Y^H \tilde{x}\|_2}{\|\tilde{x}\|_2} = \frac{\|y^H E \tilde{x}\|_2}{\|\tilde{x}\|_2} \]
\[
\frac{\|y^\top x\|_2}{\|x\|_2} \leq \frac{\|E\|_2}{\|dE\|_2} \|x\|_2 = \frac{\|E\|_2}{\text{sep}(d,M)}
\]

For \( \|x\|_2 \to 0 \) \( \tilde{x} \to x \). So,

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

So, \( \text{sep}^{-1}(d,M) \) is a condition number

For e.g. vector \( x \)

\[
\text{sep}^{-1}(d,M) \neq \text{lower bound on separation of } d \text{ from spectrum of } M.
\]

In any consistent norm,

\[
\text{sep}(d,M) \leq \min_{\mu \in \Lambda(M)} |d - \mu|
\]

\[
\text{sep}(d,M) = \|dE - M\| \geq \rho((dI - M)^{-1}) = \max_{\mu \in \Lambda(M)} |d - \mu|^{-1}
\]

\[
\text{sep}(d,M) \leq \min_{\mu \in \Lambda(M)} (d - \mu)
\]

However, \( \text{sep}(d,M) \) can be much smaller than \( \min |d - \mu| \) if an eigenvalue of \( M \) is ill-conditioned.
Hermitean case

Theo 3.16

Since eigenvalues of Hermitean matrix are perfectly well-conditioned

$$(\text{cond} \, \nu = 1)$$

we have

Let $d$ be real, $M$ Hermitean. Then in 2-norm

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

If $A$ Hermitean we can take $Y = X$

(see above), so that $M$ Herm.

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

So, it's the separation of eigenvalues that counts.
Let $X$ and $Y$ be subspaces of same dim, $X$ orthonormal basis for $X$, $Y$ for $Y$.

Then $\varphi_i = \cos \Theta_i$ are the canonical angles between $X$ and $Y$, where $\varphi_i$ are singular values of $Y^H X$.

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

$$\Theta(X, Y) = \text{diag}(\varphi_1, \ldots, \varphi_p)$$

If $Y^\perp$ orthonormal basis for $Y^\perp$ then $\Theta_i(Y^\perp, X)$ are $\sin(\varphi_i)$.

Also

$$\begin{pmatrix} Y^H \\ Y^\perp \end{pmatrix} x = \begin{pmatrix} c \\ s \end{pmatrix} \quad \text{where} \quad c^H c + s^H s = 1$$

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

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

---

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

$S_i(Z)$ nonzero singular values descending

$S_i(0)$ " " " descending

$N_i(X, Z)$ nonzero can. angles between $B(X)$ and $B(Z)$ in descending order
Note $K = U \Sigma V^H$

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

/ eig. val.s

/ eig. vec.s

Then $\sigma_i = \text{sec} \theta_i$

$\theta_i = \tan \theta_i$

$Z^H Z = (X^H + \alpha Y^H)(X + Y\alpha) = X^H X + X^H Y \alpha + \alpha^H Y^H X + \alpha^H \alpha = I + \alpha^H \alpha$

$\sigma_i = 1 + \theta_i^2$

To find canonical angles $\theta_i (X, Z)$ we need orthonormal basis for $\mathbb{R}(Z)$

Take $\hat{Z} = Z (Z^H Z)^{-\frac{1}{2}} \Rightarrow \hat{Z}^H \hat{Z} = I$

$X^H \hat{Z} = X^H (X + Y\alpha)(I + \alpha^H \alpha)^{-\frac{1}{2}}$

$= (I + \alpha^H \alpha)^{-\frac{1}{2}}$

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

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

$= 1 + \tan^2 \theta_i = 1 + \theta_i^2$

Corollary 2.5

Let $X$ be orthonormal basis for simple eig. space $X$ of $A$ and $Y$ be basis for corresponding left eig. space $X^T$

$Y^H X = I$

Then $\sigma_i (Y) = \text{sec} \theta_i (X, Y)$

and $\|Y\|_2 = \sigma_{\text{max}} (Y) = \text{sec} \theta_i (X, Y)$
Residual Analysis

Approx. eigenbasis $X$ (basis for approx. eig space):

$$R = AX - XL$$

Best $L$ ?

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

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

$$L = X^HAX \Rightarrow \text{Rayleigh quotient}$$

More generally, $X^\perp$ left inverse of $X$

Then $\exists X^\perp$ $AX$ is Rayleigh quot. of $A$

Backward Error

$X$ 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 $L = X^HAX$

then $E = (AX^H + X^H(R))$

$(A + E)X = XL$

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

Proof: by construction + taking norms
From previous
typically take
\[ L = X^H A X \]

We can proceed in 2 ways:

\[(A + \varepsilon)X = XL \text{ and consider } \]

perturbation eigenpair of perturbed matrix

Consider block deflation similarity transform and changes to \( X \) and \( L \)

\[(X \ X_L) \text{ unitary, }\]

\[
\begin{pmatrix}
X^H \\
X_L^H
\end{pmatrix}
A
\begin{pmatrix}
X \\
X_L
\end{pmatrix}
=
\begin{pmatrix}
L^H \\
G \ M
\end{pmatrix}
\]

\[A(X \ X_L) = (X_L + X_L G \ \ X_L M + X_H) \]

\[AX - XL = X_L G = R\]

\[\Rightarrow \|G\| = \|R\| \text{ for min. } \]

\[x_L^H (AX - XL) = G = X_L^H R\]

3rd. Transform:

\[
\begin{pmatrix}
I & 0 \\
-P & I
\end{pmatrix}
\begin{pmatrix}
L^H \\
G + MP - PL - PHP
\end{pmatrix}
= \]

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

\[P \text{ is l. } PL - MP = G - PHP\]

Define

\[\text{sep}(L, M) = \min \frac{\|QL - MQ\|}{\|L\|_2 = 1}\]
If \( ||G|| \leq \text{sep}^2(L, M) \) then
unique \( P \) exists s.t. \( PL-MP=G-\text{PHP} \)
and
\[
||P|| < \frac{2||G||}{\text{sep}(L, M)}
\]
Furthermore
\[
\Lambda(L+\text{PH}) \cap \Lambda(M-\text{PH}) = \emptyset
\]
Proof hinges on convergence of fixpoint iteration (solving for \( P \))
\[
G_0 = G \rightarrow \text{Solve } P_0 : P_0 L - M P_0 = G_0
\]
if \( ||G|| \) sufficiently small then \( ||P_0|| \) small
and \( ||P_0||^2 \) very small \( \rightarrow \)
\[
G_i = G_0 - P_0 \text{HP}_0 \text{ Solve } P_i L - M P_i = G_i
\]
Overall Sun. Transform for \( A \):
\[
\begin{pmatrix}
\Sigma & 0 \\
-P & \Sigma
\end{pmatrix}
\begin{pmatrix}
X \\
X_{\perp}
\end{pmatrix} = \begin{pmatrix}
E & 0 \\
0 & E
\end{pmatrix}
\begin{pmatrix}
\hat{X} \\
\hat{X}_{\perp}
\end{pmatrix}
\]
\( A \hat{X} = \hat{X}_{\perp} \) where
\( \hat{X} = X + X_{\perp} P \), \( \hat{L} = L + \text{HP} \)
From Theo 2.4 (chap4) : \( \tan \alpha_i(X, \hat{X}) = \sigma_i(P) \)
\[
\rightarrow \tan \alpha_i = \sigma_i(P) < \frac{2||G||}{\text{sep}(L, M)}
\]
Theo 2.12

\[ R = AX - XL \]
\[ S = X^TA - LXH \]

If \[ 4\|R\|\|S\| < \text{sep}^2(L,M) \]

\[ (L, \hat{X}) \text{ exists (for A)} \]

\[ A\hat{X} = \hat{X}L \]

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

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

- 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.

\[ |\text{sep}(L+E, M+F)\| - \text{sep}(L,M)\| < \|E\| + \|F\| \]

(sep defined for consistent norm)
$x^Hx = \Sigma$  \\
$\tilde{A}x = x_L + R \Rightarrow \tilde{A}x_L - x_L = R$

$(\tilde{A} - R x^H_L)x_L = x_L + R - R = x_L$ (exact)

$E = Rx^H \Rightarrow \tilde{A} = A + E$ and $Ax_L = x_L$

Further let $A = x_L y^H_1 + x_2 M y^H_2$

$\begin{pmatrix} y^H_1 \\ y^H_2 \end{pmatrix} (A + E) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} L + F_{11} & F_{12} \\ F_{21} & M + F_{22} \end{pmatrix}$

(for consistent family of norms)

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

If $\|F_{11}\| + \|F_{22}\| < S^2$ then

$P$ exists s.t. $\|P\| < \frac{2\|F_{11}\|}{S}$

$L, X_1 = (L + F_{11} + F_{12} P, x_1 + x_2 P)$

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

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