Synopsis of Numerical Linear Algebra

Eric de Sturler
Department of Mathematics, Virginia Tech
✉ sturler@vt.edu
🕸 http://www.math.vt.edu/people/sturler

Iterative Methods for Linear Systems: Basics to Research
Numerical Analysis and Software I

Systems of Linear Equations 1

$Ax = b$ $\quad A \in \mathbb{C}^{n \times n}$ $\quad b \in \mathbb{C}^n$ $\quad$ (or $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$)

A nonsingular (regular) if its column vectors are independent.

alternatives: A^{-1} exists, $\det(A) \neq 0$, $Ax \neq 0$ for all $x \neq 0$ etc

Otherwise A singular (so $\exists x \neq 0 : Ax = 0$, etc)

If A nonsingular A has unique solution for any $b \in \mathbb{C}^n$
$x = A^{-1} b$

If A singular $Ax=b$ has either no solution or infinitely many sol.s

if $Ax = b$ and $Az = 0$
then $A(x + \alpha z) = b$ for any α

In practice, matrix may be nearly singular and this may have an effect on the accuracy of an approximate solution. (more later)

Systems of Linear Equations 2

Monday, August 22, 2011
11:08 AM

For small (dense) linear systems

LU - decomposition or
Cholesky - decomp. (also factorization)
and variants

$$A = L \cdot U \quad \begin{cases} L & \text{lower triangular (unit diagonal) usually w.} \\ U & \text{upper triangular} \end{cases}$$

$$Ax = b \to LUx = b \quad \begin{cases} \text{Solve } Ly = b & \text{forward substitution} \\ \text{Solve } Ux = y & \text{backward subst.} \end{cases}$$

A HPD $\left(A = A^* \text{ and } z^* A z > 0 \text{ for } z \neq 0\right)$

$A = LL^*$ (or $A = R^* R$)

For sparse matrices (even quite large ones) special methods have been developed that reduce fill-in (to reduce complexity and storage) → sparse direct methods

For general matrices pivoting (swapping rows and columns) is often necessary:

$$PAQ = LU \quad \begin{cases} P & \text{row exchanges} \\ Q & \text{column exchanges} \end{cases}$$

(P or $Q = I$ possible)

We can also solve linear system using QR decomposition:

$$A^{n \times n} = Q^{n \times n} \cdot R^{n \times n} \quad \begin{cases} Q \text{ is unitary/orthogonal} \\ R \text{ is upper triangular} \end{cases}$$

$Q^* Q = Q Q^* = I$, Q is unitary for complex Q
$Q^T Q = Q Q^T = I$, Q is orthogonal for real Q

Other special matrices

Other special matrices

$AA^* = A^*A \rightarrow A$ normal (orthog. eigenvectors)
$A = A^* \rightarrow A$ Hermitian
$A = A^T$ (real A) $\rightarrow A$ symmetric
(complex A) $\rightarrow A$ complex symmetric

$x^*Ax > 0$ for any $x \neq 0$, A positive definite
(some people define PD only for symm/Herm matrices, others for all matrices)

Norms

A norm on a vector space V is any function $f : V \to \mathbb{R}$ such that

1. $f\left(x\right) \geq 0$ and $f\left(x\right) = 0 \Leftrightarrow x = 0$,
2. $f\left(\alpha x\right) = \left|\alpha\right| f\left(x\right)$,
3. $f\left(x + y\right) \leq f\left(x\right) + f\left(y\right)$,

where $x \in V$ and $\alpha \in \mathbb{R}$.

Important vector spaces in this course: $\mathbb{R}^n$, $\mathbb{C}^n$, and $\mathbb{R}^{m \times n}$, $\mathbb{C}^{m \times n}$ (matrices). Note that the set of all m-by-n matrices (real or complex) is a vector space.

Many matrix norms possess the submultiplicative or consistency property:

$f\left(AB\right) \leq f\left(A\right) f\left(B\right)$ for all $A \in \mathbb{C}^{m \times k}$ and $B \in \mathbb{C}^{k \times n}$ (or real matrices).

Note that strictly speaking this is a property of a *family of norms*, because in general ‘each’ f is defined on a different vector space.

Norms

We can define a matrix norm using a vector norm (an induced matrix norm):

$$\| A \|_\alpha = \max_{x \neq 0} \frac{\| Ax \|_\alpha}{\| x \|_\alpha} = \max_{\|x\|_\alpha = 1} \| Ax \|_\alpha$$

Induced norms are always consistent (satisfy consistency property).

Two norms $\| . \|_\alpha$ and $\| . \|_\beta$ are equivalent if there exist positive, real constants a and b such that

$$\forall x : a \| x \|_\alpha \leq \| x \|_\beta \leq b \| x \|_\alpha$$

The constants depend on the two norms but not on x.

All norms on a finite dimensional vector space are equivalent.

Norms

Some useful norms on $\mathbb{R}^n$, $\mathbb{C}^n$, $\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$:

p-norms: $\left\|x\right\|_p = \left(\sum_{i=1}^{n}\left|x_i\right|^p\right)^{\frac{1}{p}}$, especially $p = 1, 2, \infty$, where $\left\|x\right\|_\infty = \max_i \left|x_i\right|$.

Induced matrix p-norms are:

$\left\|A\right\|_1 = \max_j \sum_{i=1}^{n}\left|a_{ij}\right|$ (max absolute column sum)

$\left\|A\right\|_2 = \sigma_{\max}\left(A\right)$ (max singular value – harder to compute than others)

$\left\|A\right\|_\infty = \max_i \sum_{j=1}^{n}\left|a_{ij}\right|$ (max absolute row sum)

Matrix Frobenius norm:

$\left\|A\right\|_F = \left(\sum_{i,j=1}^{n}\left|a_{ij}\right|^2\right)^{\frac{1}{2}}$ (similar to vector 2-norm for a matrix)

All these norms are consistent (satisfy the submultiplicative property)

Norms & Inner Products 1

Friday, August 19, 2011
9:25 AM

Norms and Inner Products

Let V be a vector space ($\mathbb{C}^n$ or $\mathbb{R}^n$) over complex or real field

$x, y \in V \quad \alpha \in \mathbb{C}$ (or $\mathbb{R}$)

A function $f: V \to \mathbb{R}$ is a norm if

$f: V \to \mathbb{R}$ s.t.

(1) $f(x) \geq 0$ <u>and</u> $f(x) = 0 \iff x = 0$

(2) $f(\alpha x) = |\alpha| f(x)$

(3) $f(x+y) \leq f(x) + f(y)$ (triangle ineq.)

Often written $\|x\|$, $\|x\|_\alpha$, $|||x|||$

examples

$$\|x\| = \sum_{i=1}^{n} |x_i|^2$$

Important consequence of triangle ineq. :

$$\big|\|x\| - \|y\|\big| \leq \|x-y\|$$

$\|x\| = \|y + x - y\| \leq \|y\| + \|x-y\| \iff \|x\| - \|y\| \leq \|x-y\|$

same way $\|y\| - \|x\| \leq \|y - x\| = \|x-y\|$

Norms & Inner Products 2

Monday, August 22, 2011
11:26 AM

Prove induced matrix norm is consistent

Prove for any nonsingular B and any norm $\|\cdot\|_\alpha$ that $\|Bx\|_\alpha$ is a norm

Prove for B HPD (SPD in real case) that $(x^* B x)^{1/2}$ is a norm

Norms & Inner Products 3

Monday, August 22, 2011

11:32 AM

$\|Bx\|$ is norm (for) for x any nonsingular B

$\|x\|$ norm

$\|Bx\| \geq 0$ and $\|Bx\| = 0 \Rightarrow Bx = 0$

B nonsing $\rightarrow x = 0$

$$\left.\begin{array}{l} \|Bx\| = 0 \Rightarrow x = 0 \\ x = 0 \Rightarrow Bx = 0 \Rightarrow \|Bx\| = 0 \end{array}\right\} \|Bx\| = 0 \Leftrightarrow x = 0$$

$$\|B(\alpha x)\| = \|\alpha(Bx)\| = |\alpha| \, \|Bx\|$$

$$\|B(x+y)\| = \|Bx + By\| \leq \|Bx\| + \|By\|$$

$$\|A\|_\alpha = \max_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}$$ show consistent

$$\frac{\|Ay\|_\alpha}{\|y\|_\alpha} \leq \|A\|_\alpha \Leftrightarrow \|Ay\|_\alpha \leq \|A\|_\alpha \|y\|_\alpha$$

$$\|AB\|_\alpha = \max_{x \neq 0} \frac{\|ABx\|_\alpha}{\|x\|_\alpha}$$

$$\|AB\|_\alpha \leq \|A\|_\alpha \|B\|_\alpha$$ to show

$$= \left(\max_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}\right) \cdot \left(\max_{y \neq 0} \frac{\|By\|_\alpha}{\|y\|_\alpha}\right)$$

$$\max_{x \neq 0} \frac{\|ABx\|_\alpha}{\|x\|_\alpha} = \max_{x \neq 0} \frac{\|A(Bx)\|_\alpha}{\|Bx\|_\alpha} \cdot \frac{\|Bx\|_\alpha}{\|x\|_\alpha}$$

$\max \|Av\|$. $\|Bx\|_\alpha$

$$\|x\|_\alpha \qquad x\neq 0 \qquad \|Bx\|_\alpha \qquad \|x\|_\alpha$$

$$\leq \max \frac{\|Ay\|_\alpha}{\|y\|_\alpha} \cdot \max \frac{\|Bx\|_\alpha}{\|x\|_\alpha}$$

Show $(x^* B x)^{1/2}$ norm for B HPD

$x^* B x > 0$ for $x \neq 0$ (by def of PD)

$x = 0 \implies x^* B x = 0$

So, $(x^* B x)^{1/2} \geq 0$ and $(x^* B x)^{1/2} = 0 \iff x = 0$

$$\left((\alpha x)^* B (\alpha x)\right)^{1/2} = (\alpha^* \alpha)^{1/2} (x^* B x)^{1/2} = |\alpha| (x^* B x)^{1/2}$$

$$\left((x+y)^* B (x+y)\right)^{1/2} = \left(x^* B x + x^* B y + \underline{y^* B x + y^* B y}\right)^{1/2} =$$

$$\left(x^* B x + y^* B y + 2\,\mathrm{Re}\left(x^* B y\right)\right)^{1/2}$$

$$\left((x^* B x)^{1/2} + (y^* B y)^{1/2}\right)^2 = x^* B x + y^* B y + 2 (x^* B x)^{1/2} (y^* B y)^{1/2}$$

$|x^* B y| \leq (x^* B x)^{1/2} (y^* B y)^{1/2}$ → Cauchy-Schwartz

Norms & Inner Products 4

Monday, August 22, 2011
6:19 PM

We'll see many important uses of norms.
First, use to study accuracy and sensitivity of solutions.

Consider two linear systems with slightly different rhs but same matrix.

$$Ax = b \qquad \text{and} \qquad A(x+\Delta x) = b + \Delta b$$

How "large" (long) is Δx (especially relative to length of x) for small (short) Δb.

Important because

a) practical problems involve estimates, measurements, noise/error, ...

b) every computation involves some error (except symbolic but too expensive)

c) need to understand how sensitive (reliable) answer is

$$A(x+\Delta x) = \cancel{Ax} + A\Delta x = \cancel{b} + \Delta b \iff \Delta x = A^{-1}\Delta b$$

Consider any (vector) norm and its induced matrix norm

$$\|\Delta x\| = \|A^{-1}\Delta b\| \le \|A^{-1}\|\,\|\Delta b\|$$

$$\frac{\|\Delta x\|}{\|x\|} \le \|A^{-1}\| \frac{\|\Delta b\|}{\|x\|} = \|A\|\cdot\|A^{-1}\| \cdot \frac{\|\Delta b\|}{\|A\|\cdot\|x\|}$$

From $b = Ax \Rightarrow \|b\| \le \|A\|\,\|x\|$ we get

$$\frac{\|\Delta x\|}{\|x\|} \le \|A\|\,\|A^{-1}\| \cdot \frac{\|\Delta b\|}{\|b\|}$$

relative change (error) in x is bounded by

$\underbrace{\|A\|\cdot\|A^{-1}\|}$ times relative change in b

$\downarrow$

cond(A) or $\kappa(A)$ condition number of A

measures sensitivity of solution (conditioning)
depends on chosen norm.

Similar result holds more generally for changes in A and/or b

$$(A+\Delta A)(x+\Delta x) = b+\Delta b \quad \text{and} \quad Ax = b$$

$$\cancel{Ax} + A\Delta x + \Delta A x + \underbrace{\Delta A \Delta x}_{} = \cancel{b} + \Delta b \qquad (Ax=b)$$

neglect products of small factors

$$A\Delta x = \Delta b - \Delta A x \iff \Delta x = A^{-1}\Delta b - A^{-1}\Delta A x$$

$$\|\Delta x\| \le \|A^{-1}\| \, \|\Delta b\| + \|A^{-1}\| \, \|\Delta A\| \, \|x\| \qquad \iff$$

$$\frac{\|\Delta x\|}{\|x\|} \le \|A^{-1}\| \frac{\|\Delta b\|}{\|x\|} + \|A^{-1}\| \, \|\Delta A\| \qquad =$$

$$\|A\| \, \|A^{-1}\| \frac{\|\Delta b\|}{\|A\| \, \|x\|} + \|A\| \, \|A^{-1}\| \cdot \frac{\|\Delta A\|}{\|A\|} \qquad \Longrightarrow$$

$$\frac{\|\Delta x\|}{\|x\|} \le \text{cond}(A) \underbrace{\left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right)}_{\text{total relative change in linear system}}$$

Storing/computing system and any numerical computation will introduce errors of roughly machine precision ε_{mach}

double precision : $\varepsilon_{mach} \simeq 10^{-16}$

So, at best $\dfrac{\|\Delta x\|}{\|x\|} \le \text{cond}(A)\, \varepsilon_{mach}$

Norms & Inner Products 5

Monday, August 22, 2011
7:21 PM

In general, we don't know the error in the solution, becouse we don't know the exact solution.

Let $\hat{x}$ be the (unknown) exact solution to $Ax = b$
Let $\tilde{x}$ be an approximate solution

residual: $r = b - A\tilde{x}$ (easy to compute)

$r = A\hat{x} - A\tilde{x} = Ae$ (e is error $\hat{x} - \tilde{x}$)

$e = A^{-1}r \Rightarrow \|e\| \leq \|A^{-1}\| \|r\|$ so, we can bound the norm of the error directly in terms of the norm of the residual.
In practice, we often only estimate $\|A^{-1}\|$

$$\frac{\|e\|}{\|\tilde{x}\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|r\|}{\|A\| \, \|\tilde{x}\|} = \text{cond}(A) \frac{\|r\|}{\|A\| \, \|\tilde{x}\|}$$

If cond(A) not (very) large, a small relative residual implies a small relative error

Example $\text{cond}(A) = 10^6$ and $\dfrac{\|r\|}{\|A\| \, \|\tilde{x}\|} \leq 10^{-14}$

$$\frac{\|e\|}{\|\tilde{x}\|} \leq 10^{-8}$$

In single precision, we can't expect $\dfrac{\|r\|}{\|A\| \, \|\tilde{x}\|} \lesssim 10^{-8}$ (in general)
so, we cannot expect relative error to be small.

magnitude of cond nr should be considered relative to precision of machine/arithmetic and relative errors in problem (in A and in b).

precision of machine/arithmetic and relative errors in problem (in A and in b).

Norms & Inner Products 6

Monday, August 22, 2011
9:01 PM

In estimating errors we often want "backward error" bounds.

"Forward error" is usual notion of error:

difference between exact and approximate solution $(\hat{x} - \tilde{x})$
(divided by solution for relative error)

Alternatively, consider computed, approximate solution the exact solution of a perturbed problem. Then perturbation is backward error.

It is in general not unique. We want to show (in general) that a small perturbation exist such that computed answer is exact → we solved the problem with small backward error.

In general, this is the best we can expect from a good algorithm. The forward error can still be large, when the problem is sensitive. This is one of the main problems with forward error analysis, a worst case analysis for many algorithms yields a very large bound because the forward error can be large for a sensitive problem (but that doesn't mean the algorithm is bad). ~~In~~ For some algorithms cases, it is easy to see that bound is large only for special instances of the problem. For many alg.s this is, unfortunately, not the case. Backward error analysis allows us to separate the two issues. We can show our algorithm produces an exact solution for slightly perturbed problem.

Well- cond. problem → small change from solution original problem

Ill- cond. problem → possibly large change from sol. original prob.

Error Analysis

Monday, August 22, 2011
10:22 PM

Solve $Ax = b$ → approx. sol. $\tilde{x}$

Any $(\Delta A, \Delta b)$ s.t. $(A + \Delta A)\tilde{x} = b + \Delta b$ provides $\overset{\text{(backw. err.)}}{\text{B.E.}}$

We're interested in small instances.

Let $r = b - A\tilde{x}$; then $A\tilde{x} = b - r$. So, residual gives a B.E.

Small relative backw. error if $\frac{\|r\|}{\|b\|}$ is small.

Alternative $\left(A + \frac{r\tilde{x}^*}{\|\tilde{x}\|_2^2}\right)\tilde{x} = A\tilde{x} + r = b - r + r = b$ ($\tilde{x}$ exact sol.)

$$\frac{\left\| r\tilde{x}^* / \|\tilde{x}\|_2^2 \right\|}{\|A\|} = \frac{\|r\|_2 \cdot \|\tilde{x}\|_2}{\|A\| \cdot \|\tilde{x}\|_2} = \frac{\|r\|_2}{\|A\| \cdot \|\tilde{x}\|}$$

Remember $\frac{\|e\|}{\|\tilde{x}\|} \le \text{cond}(A) \cdot \frac{\|r\|}{\|A\|\,\|\tilde{x}\|}$

So, relative FE (forw. err.) bounded by cond nr. times relative BE.

This holds generally.

Note also $\frac{\|e\|}{\|x\|} \le \overset{\parallel\,\text{cond}(A)}{\kappa(A)} \cdot \frac{\|\Delta b\|}{\|b\|}$.

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 ~~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 $f(x+\varepsilon)$ (or $f(x+\varepsilon) + \underline{\underline{\Delta f}}(x+\varepsilon)$)

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

Example: Compute u^Tv $u, v \in \mathbb{R}^n$
$\hookrightarrow u_i, v_i$ machine numbers

exact answer : $\sum u_i v_i$

floating pt. $fl(u^Tv) = (u_1v_1(1+\varepsilon_1) + u_2v_2(1+\varepsilon_2))(1+\varepsilon_3) + \cdots$

usually we just write $(\ldots(u_1v_1(1+\varepsilon) + u_2v_2(1+\varepsilon))(1+\varepsilon) + u_3v_3(1+\varepsilon))(1+\varepsilon)\ldots$

and keep in mind that all "ε" are different

but $|\varepsilon| \leq \varepsilon_M$

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

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

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

$$fl(u^Tv) \cong u^Tv + u_1v_1 \overset{(\varepsilon_1+\varepsilon_2+\cdots)}{n\varepsilon} + u_2v_2 n\varepsilon + \cdots$$

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

* sign of each ε may differ → assume all errors accumulate (worst case)

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

* simplify by ~~[illegible]~~ taking $n\varepsilon|u_3v_3|$ i.s.o. $(n-1)\varepsilon|u_3v_3|$ etc

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

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

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

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

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

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

Backward error analysis:

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

$$\cancel{\simeq}\; n\varepsilon u_1 v_1 + n\varepsilon u_2 v_2 + \cdots$$

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

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

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

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

$$\tilde{u}_1 = u_1 + \eta \quad \text{where } |\eta_i| \le \tfrac{1}{2} n \varepsilon_M$$

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

So, floating pt. computation of dot product has <u>small relative backward error.</u>

What <u>explains</u> the huge relative forward error of $u^T v \simeq 0$ is the "sensitivity" or "conditioning" of the ~~[illegible]~~ 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.

Inner Products

Many methods to select z_m from the Krylov space are related to projections.

We call $f : S \times S \to \mathbb{R}$ an inner product over the real vector space S, if for all vectors x, y, z and scalars α,

1. $f(x,x) \geq 0$ and $f(x,x) = 0 \Leftrightarrow x = 0$
2. $f(\alpha x, z) = \alpha f(x,z)$
3. $f(x+y,z) = f(x,z) + f(y,z)$
4. $f(x,z) = f(z,x)$

For a complex inner product, $f : S \times S \to \mathbb{C}$, over a complex vector space S we have instead of property (4): $f(x,z) = \overline{f(z,x)}$.

Inner products are often written as $\langle x, y \rangle$, (x, y), or $\langle x, y \rangle_\alpha$, etc..

We say x and y are orthogonal (w.r.t α-IP), $x \perp_\alpha y$ if $\langle x, y \rangle_\alpha = 0$.

Inner products and Norms

Each inner product defines, or induces, a norm: $\|x\| = \sqrt{\langle x, x\rangle}$. (proof?)

Many norms are induced by inner products, but not all. Those norms that are have additional nice properties (that we'll discuss soon).
An inner product and its induced norm satisfy: $|\langle x, y\rangle| \le \|x\|\|y\|$ (CS ineq)

A norm induced by an inner product satisfies the parallelogram equality:

$$\|x+y\|^2 + \|x-y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right)$$

In this case we can find the inner product from the norm as well:

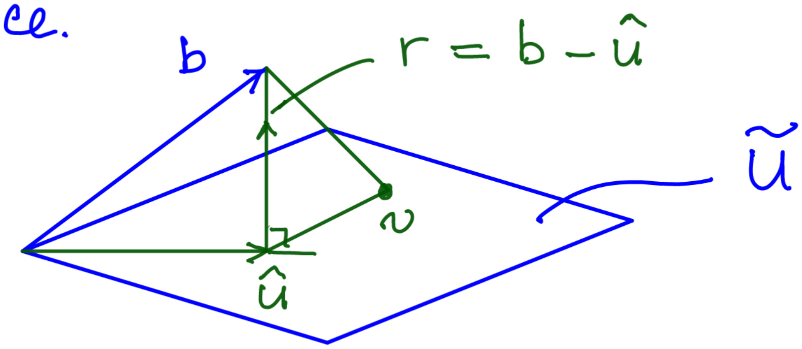
Real case: $\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$

Complex case:

$$\mathrm{Re}\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right), \quad \mathrm{Im}\langle x, y\rangle = \frac{1}{4}\left(\|x+iy\|^2 - \|x-iy\|^2\right)$$

Projections

Theo: Best approximation in inner product space.



$\|b-v\|^2 =$
$\|b-\hat{u}\|^2 + \|\hat{u}-v\|^2$
$\geq \|b-\hat{u}\|^2$

Let $\langle \cdot , \cdot \rangle$ be inner prod on $\tilde{V}$ and $\|.\|$ be its associated norm, $\|x\| = \langle x, x\rangle^{1/2}$.
Let $\tilde{U}$ be a subspace of $\tilde{V}$. Then for any $b \in \tilde{V}$ and $\hat{u} \in U$:

$$\|b-\hat{u}\| \leq \|b-v\| \text{ for all } v \in U$$

<u>iff</u> $\langle b-\hat{u}, u\rangle = 0$ for all $u \in U$.

Proof part (i)

Assume $\langle b-\hat{u}, u\rangle = 0$ for all $u \in \tilde{U}$, and $v \in U$.
Then $\|b-v\|^2 = \|b-\hat{u}+\hat{u}-v\|^2 =$
$\langle (b-\hat{u})+(\hat{u}-v), (b-\hat{u})+(\hat{u}-v)\rangle =$
$\langle b-\hat{u}, b-\hat{u}\rangle + \langle \hat{u}-v, \hat{u}-v\rangle + \langle b-\hat{u}, \hat{u}-v\rangle + \langle \hat{u}-v, b-\hat{u}\rangle$
Since $\hat{u}-v \in U$, $\langle b-\hat{u}, \hat{u}-v\rangle = \langle \hat{u}-v, b-\hat{u}\rangle = 0$.
So, $\|b-v\|^2 = \|b-\hat{u}\|^2 + \|\hat{u}-v\|^2 \geq \|b-\hat{u}\|^2$.

qed

It turns out that the orthogonality condition

$$\langle b-\hat{u}, u\rangle = 0 \text{ for all } u \in \tilde{U}$$

provides a recipe for finding $\hat{u}$ that reduces an optimization problem (hard/expensive) to solving a linear system of equations (relatively easy/cheap).

We will show this by construction, which will also show that $\hat{u}$ is unique for a given b.

part (ii)

Assume that $\|b-\hat{u}\| \leq \|b-v\|$ for all $v \in \tilde{U}$.

Now, let $z \in \tilde{U}$, $z \neq 0$, and $\langle b-\hat{u}, z\rangle = t$, and consider the points $\hat{u}+\alpha t z$ for $\alpha \in \mathbb{R}$.

(note $\hat{u}+\alpha t z \in \tilde{U}$ for any α)

Then $\|b-(\hat{u}+\alpha t z)\|^2 = \langle (b-\hat{u})-\alpha t z, (b-\hat{u})-\alpha t z\rangle$

$= \langle b-\hat{u}, b-\hat{u}\rangle - \alpha t\langle z, b-\hat{u}\rangle - \alpha\bar{t}\langle b-\hat{u}, z\rangle + \alpha^2 |t|^2 \langle z, z\rangle$

$= \|b-\hat{u}\|^2 - 2\alpha |t|^2 + \alpha^2 |t|^2 \|z\|^2$

If $|t|^2 \neq 0$, this is a parabola in α with a min less than $\|b-\hat{u}\|^2$. (just minize expression)

Hence $\|b-\hat{u}\|^2 \leq \|b-v\|^2$ for all $v \in \tilde{U}$ implies that $|t| = 0 \iff \langle b-\hat{u}, z\rangle = 0$. Since this must hold for any $z \in \tilde{U}$, $\langle b-\hat{u}, u\rangle = 0$ for all $u \in \tilde{U}$.

Next, given b we construct $\hat{u}$.

Let $u_1, \dots, u_k$ form a basis for $\tilde{U}$. Then
$\langle b-\hat{u}, u_i\rangle = 0$ for $i = 1..k$
Also $\hat{u} \in U \Rightarrow \hat{u} = \sum_{j=1}^{k} u_j \alpha_j$ (where α_j unknown)
Subst. gives:

$$\langle b - \textstyle\sum_j u_j \alpha_j , u_i \rangle = 0 \qquad i = 1..m \qquad \iff$$

$$\langle b, u_i\rangle - \textstyle\sum_j \alpha_j \langle u_j, u_i\rangle = 0 \quad \text{for } i = 1...m$$

Now let $g_{ij} = \langle u_j, u_i\rangle$, $G = (g_{ij})$, the matrix with coeff.s g_{ij} and $f_i = \langle b, u_i\rangle$

Lemma: If $\{u_1, \dots, u_k\}$ are lin. indep, then the Gram matrix G is invertible. (proof later)

Now we can write $\langle b, u_i\rangle - \sum_j \alpha_j \langle u_j, u_i\rangle = 0$

as $\sum_{j=1}^{m} g_{ij}\alpha_j = b_i$ for $i = 1..m$ or in

matrix form $G\alpha = f \Rightarrow \alpha = G^{-1}f$ and

$\hat{u} = \sum_{j=1}^{k} u_j \alpha_j$.

(note that f depends linearly on b and α depends linearly on f. Hence, $\hat{u}$ depends linearly on b)

Best approx. in $\mathbb{C}^m$ with std in.prod / norm

Let $\tilde{U}$ be subspace of $\mathbb{C}^m$ and $\{u_1, \dots, u_k\}$ be basis for $\tilde{U}$.

Then $u_i \in \mathbb{C}^m \rightarrow$ matrix $U = (u_1 \cdots u_k)$

$R(U) = \tilde{U}$ and $\hat{u} = U\alpha$ for some $\alpha \in \mathbb{C}^k$

Using std. inner product $g_{ij} = \langle u_j, u_i \rangle = u_i^* u_j$ and $G = U^*U$

The orthog. cond. $b - U\alpha \perp \tilde{U} \Rightarrow$

$b - U\alpha \perp u_i$ for $i = 1..k \rightarrow U^*(b - U\alpha) = 0$

$\Leftrightarrow U^*b - U^*U\alpha = 0 \rightarrow G\alpha = f$ with

$G = U^*U$ and $f = U^*b$

Again $\alpha = G^{-1}f = (U^*U)^{-1}U^*b$ and

$\hat{u} = U\alpha = U(U^*U)^{-1}U^*b$

Proof of Lemma for Gram matrix for this case:

U^*U is invertible <u>iff</u> lin. indep.

Assume $\{u_1 \dots u_k\}$ is lin. indep.

Let $U^*Uz = 0$. Then $z^*U^*Uz = 0 \Leftrightarrow$ $\|Uz\|_2^2 = 0$, which implies $Uz = 0$. Since the columns of U are lin. indep. $z = 0$ must hold.

Hence U^*U is invertible.

Conversely, assume U^*U is invertible.

Let $\sum_{j=1}^{k} u_j z_j = Uz = 0$. Then $U^*Uz = U^*0 = 0$ and $z = (U^*U)^{-1}(U^*U)z = 0$.

The same arguments apply for a general inner product.

We saw that the best approx. problem to b in the space $\tilde{U} \subseteq \mathbb{C}^m$ was solved by $\hat{u} = U(U^*U)^{-1}U^*b$

Let $P = U(U^*U)^{-1}U^*$. Verify that $P^2 = P$. Such a matrix or, more generally, linear transformation is called a projection (or projector).

Def: let $P : \tilde{V} \to \tilde{V}$ be a linear transf. and $P^2 = P$. Then P is a projection.

The particular choice for P above satisfies another important property.

Def: Let $\tilde{V}$ be a vector space w. inner product $\langle \cdot, \cdot \rangle_{\tilde{V}}$. A linear transf. $T: \tilde{V} \to \tilde{V}$ (or linear oper.) is called selfadjoint if

$$\forall x, y \in \tilde{V} : \langle Tx, y \rangle_{\tilde{V}} = \langle x, Ty \rangle_{\tilde{V}}$$

In $\mathbb{C}^m$ with the std inner product $(\langle x, y \rangle = y^* x)$ a matrix is self-adjoint if it is Hermitian:

$$A \in \mathbb{C}^{m \times m} : \quad \langle Ax, y \rangle = \langle x, Ay \rangle \iff$$
$$y^* A x = y^* A^* x$$

Taking $x, y = e_i, e_j$ (Cartesian basis vectors) for $i, j = 1..m$ shows this component-wise.

$P = U(U^* U)^{-1} U^*$ is self-adj. wrt the std in.prod.

Def: An orthog. proj. is a lin. transf. $P: \tilde{V} \to \tilde{V}$ such that $P = P^2$ and P is self-adjoint wrt $\langle \cdot, \cdot \rangle_{\tilde{V}}$.

An orthog. proj. P has the property that $\| b - Pb \|_V \leq \| b - u \|_V$ for all $u \in R(P)$ (already proved above)

Def: Let $\tilde{U}, \tilde{V} \subseteq \tilde{W}$ ($\tilde{W}$ vector space)

$$\tilde{U} + \tilde{V} = \{ u + v \mid u \in \tilde{U}, v \in \tilde{V} \}$$

Def: Let $\tilde{U}, \tilde{V}$ be subspaces of $\tilde{W}$ (vector space)

We say that $\tilde{U}$ and $\tilde{V}$ form a direct sum decomposition of $\tilde{W}$: $\tilde{U} \oplus \tilde{V} = \tilde{W}$ if

1) $\tilde{U} + \tilde{V} = \tilde{W}$ (or $\forall w \in \tilde{W} : w = u + v$, where $u \in \tilde{U}$, $v \in \tilde{V}$)

2) $\tilde{U} \cap \tilde{V} = \{\underline{0}\}$

(note two subspaces $\tilde{U}, \tilde{V}$ form a direct sum of $\tilde{U} \oplus \tilde{V}$ if $\tilde{U} \cap \tilde{V} = \{\underline{0}\}$.)

Properties of Projections

Let $P : \tilde{V} \to \tilde{V}$ be projection. Then $(I-P)$ is projection (called complementary projection)

$$(I-P)^2 = I - P - P + P^2 = I - P$$

$R(P) = N(I-P)$, $N(P) = R(I-P)$ (verify)

Theo: $R(P) \oplus N(P) = \tilde{V}$

Proof: Let $v \in \tilde{V}$. Then $v = v - Pv + Pv =$

$$(I-P)v + Pv, \text{ where } \begin{array}{l} (I-P)v \in N(P) \\ Pv \in R(P) \end{array}$$

So, $R(P) + N(P) = \tilde{V}$

Let $x \in R(P)$ and $x \in N(P)$. Then $x = Py$, for some y, and $Px = 0$. So, $\underline{0} = Px = P^2y = Py = x$.

Hence, $R(P) \cap N(P) = \{\underline{0}\}$.

Note that $x \in R(P) \Rightarrow Px = x$, so x eig. vec. with eig. val 1. Also, $x \in N(P) \Rightarrow Px = 0$, so x is eig. vec with eig. val. 1.

Since every vector $v \in \tilde{V}$ can be written (uniquely) as $v = x + y$, where $x \in R(P)$ and $y \in N(P)$, the union of a basis for $R(P)$ and a basis for $N(P)$ is a basis for V (this holds generally for direct sums). Hence P is diagonalizable and has eigenvalues $\Lambda(P) = \{0, 1\}$.

So, if $P: \mathbb{C}^m \to \mathbb{C}^m$ and $\{r_1, \dots, r_k\}$ basis for $R(P)$, $\{n_1, \dots, n_\ell\}$ basis for $N(P)$, then

$$P(r_1 \dots r_k \; n_1 \dots n_\ell) = (r_1 \dots r_k \; n_1 \dots n_\ell)\begin{pmatrix} I_k & \\ & 0_\ell \end{pmatrix}.$$

For $P = U(U^*U)^{-1}U^*$, $R(P) = R(U) = \tilde{U}$.

$N(P)$?

Since U has lin. indep. col.s (basis for $\tilde{U}$)

$Py = 0$ iff $U^*y = 0 \Rightarrow u_1^*y = 0 \dots u_k^*y = 0$

that is $\langle y, u_j \rangle = 0$ for $j = 1 \dots k \Rightarrow$
$\langle y, u \rangle = 0$ for all $u \in \tilde{U}$

The set of all such y is called the orthog. complement of $\tilde{U}$:

$$\tilde{U}^\perp = \{ y \in \mathbb{C}^m \mid \langle y, u \rangle = 0 \text{ for all } u \in \tilde{U} \}$$

So, $N(P) = \tilde{U}^\perp$.

So, a projection always defines a direct sum decomposition:

$$R(P) \oplus N(P) = V$$

but not necessarily $N(P) = R(P)^\perp$. The latter holds only for orthogonal proj.s

It turns out that, conversely, every direct sum defines a projection.

Let $\tilde{U} \oplus \tilde{V} = \tilde{W}$.

<u>Theo</u>: The sum $w = u + v$, with $u \in \tilde{U}$, $v \in \tilde{V}$ (for any $w \in \tilde{W}$) is unique.

Pr. Let $w \in \tilde{W}$ and $w = u_1 + v_1$, $u_1 \in \tilde{U}, v_1 \in \tilde{V}$

and $w = u_2 + v_2$ $u_2 \in \tilde{U}, v_2 \in \tilde{V}$

Then $u_1 + v_1 = u_2 + v_2 \Leftrightarrow u_1 - u_2 = v_2 - v_1$

Since $\tilde{U}, \tilde{V}$ are vector spaces $u_1 - u_2 \in \tilde{U}$ and $v_2 - v_1 \in \tilde{V}$. But then $\tilde{U} \cap \tilde{V} = \{\underline{0}\}$ (from def. of direct sum) implies

$$\left.\begin{array}{l} u_1 - u_2 = 0 \Rightarrow u_1 = u_2 \\ V_2 - V_1 = 0 \Rightarrow V_1 = V_2 \end{array}\right\} \text{Hence the } u, v \text{ in } w = u + v \text{ are unique.}$$

So, following discussion of projections above define P such that $R(P) = \tilde{U}$ and $N(P) = \tilde{V}$ and $P^2 = P$.

Let $w \in \tilde{W}$ and $w = u + v$ $(u \in \tilde{U}, v \in \tilde{V})$

$N(P) = \tilde{V} \Rightarrow Pw = Pu + \underbrace{\cancel{Pv}}_{=0} = Pu$

$P^2 = P \Rightarrow Pu = u$

$(u \in R(P) \Rightarrow u = Px$, for some x

$\Rightarrow Pu = P^2 x = Px = u)$

So $Pw = u$ (verify that P is linear)

$\hookrightarrow$ the $\tilde{U}$ component of w.

Note that for $v \in \tilde{V}$: $v = \underset{\in \tilde{U}}{\underline{0}} + \underset{\in \tilde{V}}{v}$ and

$P v = \underline{0} \Rightarrow \tilde{V} \subseteq N(P)$

Conversely, if $x \in N(P)$, then $x = \underline{0} + x_V$, $x_V \in \tilde{V}$ and $x \in \tilde{V}$. So, $N(P) \subseteq \tilde{V}$. Hence, $N(P) = \tilde{V}$.

In practice to find the $\tilde{U}$ comp. and $\tilde{V}$ comp. (so, we can pick the $\tilde{u}$ comp.) would be expensive - essentially a large linear solve. So, we define projections using an auxiliary space (or basis for such a space) and the complementary space $\tilde{V}$ is defined implicitly.

Example: Find $\hat{u} \in \tilde{U}$ s.t. $b - \hat{u} \perp \tilde{Y}$

If $\dim \tilde{U} = k$ we need in principle $\dim \tilde{Y} = k$:

$b - \sum u_j \alpha_j \perp y_1 \dots y_k \Rightarrow$ (std in. prod.)

$$Y^*(b - U\alpha) = 0 \iff Y^* b = Y^* U \alpha$$

α uniquely defined for any b if $Y^* U$ invertible

Let $Pb = U\alpha \Rightarrow R(P) = \tilde{U}$.

What is $\tilde{V} = N(P)$?

$Pb = \underline{0} \Rightarrow \alpha = 0 \Rightarrow \alpha = (Y^*U)^{-1} Y^* b = 0$
$\Rightarrow b \in R(Y)^{\perp}$. So, $\tilde{V} = R(Y)^{\perp}$.
$Pb = U(Y^*U)^{-1} Y^* b \rightarrow P = U(Y^*U)^{-1} Y^*$

This defines a projection if Y^*U invertible which is equivalent to $R(Y)^{\perp} \cap R(U) = \{\underline{0}\}$.

As we compute $\tilde{U}$ (or its basis) and $\tilde{Y}$ (or its basis) simultaneously, at
the projection may not be defined and the method breaks down.

If $\tilde{Y} = \tilde{U}$, so $\tilde{V} = \tilde{Y}^{\perp}$, P is an orthogonal projection (wrt the std inner product).

Approximations from subspaces

If we pick an approx. from a subspace and set the residual orthogonal to that subspace, we often call the approx. a (Ritz) Galerkin approx. (and the cond. a Galerkin cond.)

For example, solving approx. $Ax = b$ using $\tilde{u}$ subspace of $\mathbb{C}^m$ to find approx.:

Find $\hat{x} \in \tilde{U}$ s.t. $b - A\hat{x} \perp \tilde{U}$ (in some inner product).

If $R(U) = \tilde{U}$: $b - AU\xi \perp U$

std in.prod. gives $U^*(b - AU\xi) = 0$.

If we use an additional space $\tilde{V}$ to define the projection, we call the approx. a Petrov-Galerkin approx. (Petrov-Galerkin condition).

Find $\hat{x} \in \tilde{U}$ s.t. $b - A\hat{x} \perp \tilde{V}$ $(\dim \tilde{V} = \dim \tilde{U})$

If $R(V) = \tilde{V}, R(U) = \tilde{U}$ and std. in.prod. used

$$V^*(b - AU\xi) = 0$$

Least Squares 1

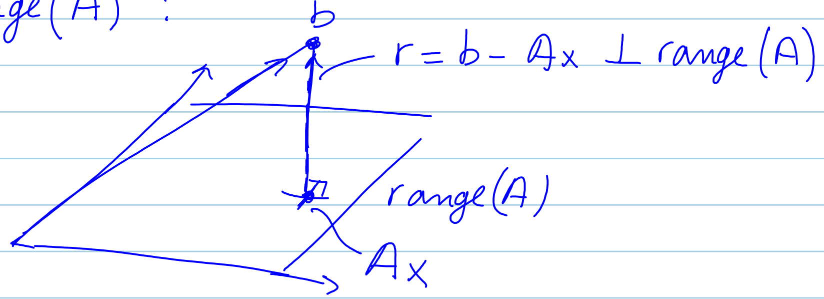
Tuesday, August 30, 2011
9:39 PM

$A \in \mathbb{C}^{n\times m}$ $n \geq m$
$b \in \mathbb{C}^{n}$

$\min\limits_{x\in\mathbb{C}^m} \|b - Ax\|_2$ (also written $Ax \approx b$)

If $b \in \text{range}(A)$ then $\exists x : Ax = b$ (by def.)

What if $b \notin \text{range}(A)$?



$A^*r = A^*(b - Ax) = 0 \iff A^*b - A^*A\,x = 0$

slight simplification : col.s A independent.

Then A^*A nonsingular (why?) : $x = (A^*A)^{-1}A^*b$

$Ax = A(A^*A)^{-1}A^*b$ $b - Ax = (I - A(A^*A)^{-1}A^*)b$

Let $P = A(A^*A)^{-1}A^*b$ $\begin{cases} P^2 = A\cancel{(A^*A)^{-1}}\,\cancel{A^*A}(A^*A)^{-1}A^* = P \\ \qquad\qquad \hookrightarrow \text{P projection} \\ P = P^* \text{ Herm/Symm (if real)} \end{cases}$

$P^2 = P$ and $P = P^*$ → P orthogonal projection
(since $\text{range}(I-P) \perp \text{range}(P)$)

$P^*(I-P) = P^* - P^*P = P - P^2 = P - P = 0$

$b = Ib = (P + (I-P))b = Pb + (I-P)b = b_1 + b_2$ $\begin{cases} b_1 \in \text{range}(A) = \text{range}(P) \\ b_2 \perp \text{range}(A) \end{cases}$

Solve $Ax = b_1$ $r = (I-P)b = b_2$

Least Squares 2

Tuesday, August 30, 2011
9:57 PM

Theo: let $x = \arg\min_{\tilde{x} \in \mathbb{C}^m} \|b - A\tilde{x}\|_2$.

Then $b - Ax \perp \text{range}(A)$

Proof: Consider $f(x) = \langle b - Ax, b - Ax \rangle$ (quadr. function)

minimum for $\hat{x}$ ($f(\hat{x})$ minimal)

Then for <u>any</u> $p \in \mathbb{C}^m \neq 0$ we must have

$$\frac{d}{dp} f(\hat{x}) = 0 \rightarrow \lim_{\substack{t \to 0 \\ (t \in \mathbb{R})}} \frac{f(\hat{x}+tp) - f(\hat{x})}{t} = 0 \quad (\text{any } p \in \mathbb{C}^m)$$

$$\frac{\langle b - A(\hat{x}+tp), b - A(\hat{x}+tp) \rangle - \langle b - A\hat{x}, b - A\hat{x} \rangle}{t} =$$

$$\frac{1}{t}\left(\cancel{\langle b - A\hat{x}, b - A\hat{x} \rangle} - \langle b - A\hat{x}, Atp \rangle - \langle Atp, b - A\hat{x} \rangle + \langle Atp, Atp \rangle - \cancel{\langle b - A\hat{x}, b - A\hat{x} \rangle}\right) =$$

$$\frac{1}{t}\left(-t\langle b - A\hat{x}, Ap \rangle - t\langle Ap, b - A\hat{x} \rangle + t^2\langle Ap, Ap \rangle\right) =$$

$$-\langle b - A\hat{x}, Ap \rangle - \langle Ap, b - A\hat{x} \rangle + t\langle Ap, Ap \rangle$$

$$\lim_{t \to 0} -\langle b - A\hat{x}, Ap \rangle - \langle Ap, b - A\hat{x} \rangle + t\langle Ap, Ap \rangle = 0 \quad (\text{any } p) \iff$$

(why?)

$$\langle b - A\hat{x}, Ap \rangle + \langle b - A\hat{x}, Ap \rangle = 0 \quad (\text{any } p) \iff$$

$$\langle b - A\hat{x}, Ap \rangle = 0 \quad (\text{any } p)$$

$b - A\hat{x} \perp Ap$ for any $p \iff b - A\hat{x} \perp \text{range}(A)$

proj. / orthog. proj.

$P^2 = P$ and not $P^* = P$ → oblique / skew proj

U = range (P) → $P : \mathbb{C}^n \to U$

$P^2 = P \Rightarrow P_U = I_U$ (restr. of P to U)

any $u \in U$: $\exists z$ s.t. $Pz = u$

then $Pu = P(Pz) = P^2 z = Pz = u$

→ any proj.

V = range $(I - P)$ $\qquad P^*(I-P) = P^* - P^*P \neq 0$ (in gen)

$v \in V \to v = (I-P)z$

$$Pv = (P - P^2)z = 0$$

Also. if $Pv = 0 \Rightarrow (I-P)v = v$

$v \in$ range $(I-P)$ (also proj. so —)

So $v \in$ null$(P) \Rightarrow v \in$ range$(I-P)$ | and

$v \in$ range$(I-P) \Rightarrow v \in$ null(P) | equal

$z \neq 0 \quad z \in$ range$(P) \Rightarrow z \notin$ null(P)

$z \in$ null$(P) \Rightarrow z \notin$ range(P)

disjound.

any $z = z_1 + z_2$ s.t.

$z_1 \in$ range(P) and $z_2 \in$ null$(P) =$ range$(I-P)$

→ direct sum $\left\{ \begin{array}{l} z = Iz = (P + (I-P))z = \\ Pz + (I-P)z = \\ z_1 + z_2 \end{array} \right.$

given U, V as above → any z: $\underbrace{\hat{U}, \hat{V}}$

range$(\hat{U}) = U$
range$(\hat{V}) = V$

$$z = \hat{U} y_1 + \hat{V} y_2$$

$$P z = U y_1 \qquad (I - P) z = V y_2$$

$$P^2 z = P U y_1 = U y_1 \quad - -$$

$$\left\{ \begin{aligned} P &= [\hat{U} \ \hat{V}] \begin{bmatrix} I & \\ & 0 \end{bmatrix} [\hat{U} \ \hat{V}]^{-1} \\ I - P &= [\hat{U} \ \hat{V}] \begin{bmatrix} 0 & \\ & I \end{bmatrix} [\hat{U} \hat{V}]^{-1} \end{aligned} \right.$$

Least Squares 3

Tuesday, August 30, 2011
10:16 PM

Solving least squares problems: $A = [a_1 \cdots a_m]$

1) (Modified) Gram-Schmidt

① $q_1 = a_1/\|a_1\|_2$ and $r_{11} = \|a_1\|_2 \to a_1 = q_1 r_{11}$

② $r_{12} = q_1^* a_2$, $r_{22} = \|a_2 - q_1 r_{12}\|_2$

$q_2 = (a_2 - q_1 r_{12})/r_{22} \iff r_{22} q_2 = a_2 - q_1 r_{12} \iff a_2 = [q_1\ q_2]\begin{bmatrix} r_{12} \\ r_{22} \end{bmatrix}$

$[a_1\ a_2] = [q_1\ q_2]\begin{bmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{bmatrix}$

note $q_i^* q_i = 1$ $(i = 1,2)$ and

$q_1^* q_2 = (q_1^* a_2 - \cancel{q_1^*} q_1 q_1^* a_2)\, r_{22}^{-1} = 0 \to q_2 \perp q_1$

ⓚ $r_{ik} = q_i^* a_k$ $i = 1..k-1$ $r_{kk} = \|a_k - \sum_{i=1}^{k-1} q_i r_{ik}\|_2$

$q_k = (a_k - q_1 r_{1k} - q_2 r_{2k} - \cdots - q_{k-1} r_{k-1k})\, r_{kk}^{-1}$

$q_i^* q_k = q_i^* a_k - q_i^* q_i r_{ik} = 0$ $(q_i^* q_j = \delta_{ij})$

↳ prove by induction

similarly prove by induction $\mathrm{range}([a_1 \cdots a_k]) = \mathrm{range}([q_1 \cdots q_k])$

$A^{n\times m} = Q^{n\times m} R^{m\times m}$ where

$Q^* Q = I_m$ and $R = \begin{pmatrix} r_{11} & & \\ & \ddots & \\ 0 & & r_{mm} \end{pmatrix}$

If $\{a_1, \dots, a_m\}$ indep → $\{q_1, \dots, q_m\}$ indep and R nonsingular

(If col.s A dependent we can still generate

$AP = Q^{n\times m} R^{m\times m}$ with $Q^* Q = I$ and

$\downarrow$ $R = (r_{11}$

$A P = Q R$ with $Q\ Q = I$ and

$\downarrow$

column exchanges
$$R = \begin{pmatrix} r_{11} & & \\ & r_{pp} & - r_{pm} \\ 0 & & 0 \end{pmatrix}$$

with slightly adapted algorithm

For accuracy compute $r_{ik} = q_i^* (a_k - \sum_{j=1}^{i-1} q_j r_{jk})$ $i = 1..m-1$

also if $\| a_k - \sum_{j=1}^{k-1} q_j r_{jk} \| \ll \| a_k \|$ then likely

compute q_k inaccurate (not $\perp q_1 \cdots q_{k-1}$)

may need reorthogonalization if $Q^* Q = I$ important

Assume A indep cols

Since range(Q) = range(A) and $Q^* Q = I_m$:

$QQ^* = A(A^*A)^{-1}A^*$ (verify)

$QQ^*b = b_1$ $(I - QQ^*)b = b_2$

$Q^*b - Q^*Ax = 0 \iff Q^*b - Rx = 0$ (solve)

$b - Ax = r + QQ^*b - QRx = r + Q(Q^*b - Rx)$

$\|b - Ax\|_2^2 = \|r\|_2^2 + \|Q^*b - Rx\|_2^2$

$\|Qz\|_2 = (z^* \cancel{Q^*Q} z)^{1/2} = (z^*z)^{1/2} = \|z\|_2$

$\uparrow$ does not change length
or angles $(Qz)^*(Qy) = z^*y$

Householder Transformation (reflection)

(1) $$z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \in \mathbb{C}^n \qquad Hz = \begin{pmatrix} \xi \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

where $|\xi| = \|z\|_2$

$$H = I - 2\frac{vv^*}{v^*v} \quad \text{or} \quad H = I - 2vv^*$$

with $\|v\|_2 = 1$

$$H^{\cancel{*}*}H = H^2 = \left(I - 2\frac{vv^*}{v^*v}\right)\left(I - 2\frac{vv^*}{v^*v}\right) =$$

$$I - 4\frac{vv^*}{v^*v} + 4\frac{v\cancel{v^*v}v^*}{(v^*v)^{\cancel{2}}} = I$$

Since $H = H^*$, obviously $HH^* = I$
$\rightarrow$ H unitary (orthog. if real)

How to pick v for (1) (fix ξ later)

(assume $\|v\| = 1$)

$$Hz = z - 2vv^*z = \xi e_1 \implies$$

$$v(2 \cdot v^*z) = z - \xi e_1$$

want v normalized so scalar $2v^*z$ doesn't matter (doesn't change direction)
pick $v = z - \xi e_1$ and normalize
(assuming $z \neq \xi e_1 \rightarrow v = 0$)

Pick ξ to avoid poss. cancellation

$$\xi = -\text{sign}(z_1)$$

QR decomp.

$$H_1^{*} A = \begin{pmatrix} \alpha_1 & \tilde{a}_{12} & \tilde{a}_{13} & \cdots \\ 0 & \tilde{a}_{22} & \vdots & \\ | & | & \vdots & \\ 0 & & & \end{pmatrix}$$ $\tilde{a}_{ij} \to$ changed

H_2 s.t. $$H_2(H_1 A) = \begin{pmatrix} \alpha_1 & \tilde{a}_{12} & x & - & x \\ 0 & \alpha_2 & | & & | \\ | & 0 & | & & | \\ | & | & | & & | \\ 0 & 0 & x & - & x \end{pmatrix}$$

$\downarrow$

~~to~~ touch only row 2 and below

$\hat{H}_2^{(n-1)\times(n-1)}$ s.t. $$\hat{H}_2 \begin{pmatrix} \tilde{a}_{22} \\ \vdots \\ \check{a}_{n2} \end{pmatrix} = \begin{pmatrix} \alpha_2 \\ 0 \\ | \\ 0 \end{pmatrix}$$

(as before)

$$H_2 = \left(\begin{array}{c|c} 1 & \\ \hline & \hat{H}_2 \end{array}\right)$$ etc

$H_i \in \mathbb{C}^{n\times n}$ $$H_n H_{n-1} H_{n-2} \cdots H_1 A = \begin{pmatrix} \alpha_1 & - & \\ & \ddots & | \\ & & \alpha_m \\ 0 & - & 0 \end{pmatrix} \equiv R$$

$$A = \overline{H_1 H_2 \cdots H_{n-1}} H_n R = QR$$

$$A^{n\times m} = Q^{n\times n} R^{n\times m} = Q^{n\times n} \begin{bmatrix} R_m \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \overset{n\times m}{Q_1} & \overset{n\times(n-m)}{Q_2} \end{bmatrix} \begin{bmatrix} R_m \\ 0 \end{bmatrix} = Q_1 R_m = Q_m R_m$$

$r(Q_1) = r(A)$ $\quad r(Q_2) = r(A)^{\perp} = null(A^*)$

$$\|b - Ax\|_2 = \|QQ^*b - Q^*Q^*Ax\|_2$$

$$= \left\|Q\left(Q^*b - \begin{bmatrix} R_m \\ 0 \end{bmatrix} x\right)\right\|_2 =$$

$$\left\|\begin{pmatrix} Q_1^*b \\ Q_2^*b \end{pmatrix} - \begin{pmatrix} R_m \\ 0 \end{pmatrix} x\right\|_2 \quad \cancel{2}$$

Solve $Q_1^*b = R_m x$

$$\|b - Ax\|_2 = \|Q_2^*b\|_2$$

$$r = Q_2Q_2^*b \qquad Q_1Q_1^*b = Ax$$

---

cols of A not indep:

$$AP = \begin{array}{c} \begin{array}{ccc} m_1 & m_2 & n-m \end{array} \\ \begin{bmatrix} Q_1 & Q_2 & Q_3 \end{bmatrix} \end{array} \begin{array}{cc} R_1 & R_{p2} \\ \left[\begin{array}{cc} \triangledown & \square \\ 0 & 0 \\ 0 & 0 \end{array}\right] \end{array} \begin{array}{l} \\ m_1 \\ m_2 \\ \cancel{m_3}\ n-m \end{array}$$

$$m = m_1 + m_2$$

---

avoidigng overflow in computing α

~~also~~ $$s = \frac{b}{(|a|^2+|b|^2)^{1/2}} = \frac{b/|b|}{\left(\frac{|a|^2}{|b|^2}+1\right)^{1/2}}$$

$$\uparrow \quad |b| \geq |a|$$

$$= \frac{b/|a|}{\left(1+\frac{|b|^2}{|a|^2}\right)^{1/2}|a|} \quad \swarrow \quad |a| > |b|$$

Same for computing ~~α~~ c

$$\tau = \frac{|a|}{|b|} \rightarrow s = \frac{b/|b|}{(\tau^2+1)^{1/2}}$$

<u>or</u>

$$\tau = \frac{|b|}{|a|} \rightarrow s = \frac{b/|a|}{(1+\tau^2)^{1/2}}$$

same for c

Can use Givens rotations to selectively set coeff. in vector (or matrix) to zero

$$\begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & & \bar{s} & & \\ & & & \ddots & & & \\ & & -s & & \bar{c} & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix} \begin{pmatrix} \vdots \\ a_i \\ \vdots \\ a_j \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \alpha \\ \vdots \\ 0 \\ \vdots \end{pmatrix}$$

Givens Transformations (rotations)

real $G = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$

complex $G = \begin{pmatrix} c & \bar{s} \\ -s & \bar{c} \end{pmatrix}$ or $\begin{pmatrix} c & \bar{s} \\ -s & c \end{pmatrix}$ if we pick $c \in \mathbb{R}$ (below)

verify $\begin{cases} G^*G = GG^* = I & \text{(complex)} \\ G^TG = GG^T = I & \text{(real)} \end{cases}$

$$\begin{pmatrix} c & \bar{s} \\ -s & \bar{c} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix}$$ where $\alpha = \left\| \begin{pmatrix} a \\ b \end{pmatrix} \right\|_2$ (or $|\alpha| = \| - \|_2$)

$$\begin{cases} ca + \bar{s}b = \alpha \\ -sa + \bar{c}b = 0 \end{cases}$$ (or $\alpha = \left\| \begin{pmatrix} a \\ b \end{pmatrix} \right\|_2 \frac{a}{|a|}$)

→ if $a = 0 \rightarrow \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

$b = 0 \rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$$\begin{cases} sca + |s|^2 b = s\alpha \\ -sca + |c|^2 b = 0 \end{cases} \rightarrow b = s\alpha \Rightarrow s = b/\alpha$$

(avoids underflow & overflow)

$$-\frac{ab}{\alpha} + \bar{c}b = 0 \Rightarrow \bar{c} = \frac{a}{\alpha} \rightarrow c = \frac{\bar{a}}{\alpha}$$

taking $\alpha = \left\| \begin{pmatrix} a \\ b \end{pmatrix} \right\|_2 \frac{a}{|a|} \rightarrow \bar{c} = \cancel{a} \cdot \frac{|a|}{\cancel{a}} \left\| \begin{pmatrix} a \\ b \end{pmatrix} \right\|_2^{-1} \in \mathbb{R}$

$$A^{n\times m} = Q^{n\times n} R^{n\times m} = \hat{Q}^{n\times m} \hat{R}^{m\times m}$$

$$Q^*Q = QQ^* = I \qquad R = \begin{pmatrix} \overset{m}{\nabla} \\ 0 \end{pmatrix} \begin{matrix}]\,m \\ ]\,n-m \end{matrix}$$

$\hat{Q}^{n\times m} \hat{R}^{m\times m}$ $\left\{ \begin{array}{l} \text{economy } QR \\ \text{thin } QR \\ \dots \end{array} \right.$

If columns of A not indep → $A = \hat{Q}^{n\times \tilde{m}} \hat{R}^{\tilde{m}\times \tilde{m}}$

→ QR decomp w. pivoting (in general robust but not fail-safe)

$$\| Q^* Ax - Q^* b \| = \| Ax - b \|_2$$

$$\left(\| Q^* x \| = \left(x^* \underline{QQ^*} x \right)^{1/2} = \|x\| \quad \text{any } x \right.$$

$$\downarrow$$

$$\left\| \begin{pmatrix} \nabla \\ \hline 0 \end{pmatrix} x = \begin{pmatrix} \overset{B_1}{\cancel{b_1}} \\ \hline \underset{B_2}{\cancel{b_2}} \end{pmatrix} \right\|$$

if col. A indep → R nonsing

solve $Rx = \cancel{b_1} B_1$

nothing to be done about $\overset{B_2}{\cancel{b_2}}$ part

$$\min \| Ax - b \|_2 = \min \left\| \begin{pmatrix} \nabla \\ 0 \end{pmatrix} x - \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\|_2 = \| b_2 \|_2$$

$$b_1 \in R(A) \to Ax = b_1 \qquad b_2 \perp R(A)$$

$$Q = [\hat{Q}_1 \; Q_2] \qquad \underset{n\times m}{A} = Q_1 \underset{m\times m}{R_1} \to \underset{n\times n}{\begin{pmatrix} Q_1^* \\ Q_2^* \end{pmatrix}} Q_1 R_1 =$$

$$Q_1 = \hat{Q} \qquad \begin{pmatrix} I \\ 0 \end{pmatrix} R_1 \; \cancel{\begin{pmatrix} R_1 \\ 0 \end{pmatrix}} = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$

$$Q_1 Q_1^* = A(A^*A)^{-1}A^*$$

$$\rightarrow Q_1 R_1 \left(R_1^* \cancel{Q_1^*} Q_1 R_1\right)^{-1} R_1^* Q_1^*$$

$$= Q_1^* R_1 \left(R_1^* R_1\right)^{-1} R_1^* Q_1^*$$

$$= Q_1^* \cancel{R_1 R_1^{-1}} \, \cancel{R_1^{-*} R_1^*} Q_1^* \quad = Q_1 Q_1^*$$

$$\mathcal{R}(Q_1) = \mathcal{R}(A)$$

$$\mathcal{R}(Q_2) = \mathcal{R}(A)^{\perp}$$

Eigenvalues and Eigenvectors

Let $Ax = \lambda x$ and $y^* A = \lambda y^*$ (for same λ).

We call the vector x a (right) eigenvector, the vector y a left eigenvector, and λ an eigenvalue of A, the triple together is called an eigentriple (of A), and (λ, x) and (λ, y) a (right) eigenpair and left eigenpair.

The set of all eigenvalues of A, $\Lambda(A)$, is called the spectrum of A (when convenient we will count multiplicities in $\Lambda(A)$).

If the matrix A is diagonalizable (has a complete set of eigenvectors) we have

$$A = V \Lambda V^{-1} \Leftrightarrow AV = V\Lambda,$$

where V is a matrix with the right eigenvectors as columns and Λ is a diagonal matrix with the eigenvalues as coefficients. This is not always possible (soon).

A similar decomposition can be given for the left eigenvectors.

Spectral Radius

The spectral radius $\rho\left(A\right)$ is defined as $\rho\left(A\right) = \max\left\{\left|\lambda\right| : \lambda \in \Lambda\left(A\right)\right\}$.

Theorem:
For all A and $\varepsilon > 0$ a consistent norm $\left\|.\right\|_{\alpha}$ exists such that $\left\|A\right\|_{\alpha} \leq \rho\left(A\right) + \varepsilon$.

So, if $\rho\left(A\right) < 1$, then a consistent norm $\left\|.\right\|_{\alpha}$ exists such that $\left\|A\right\|_{\alpha} < 1$.
Take $\varepsilon = {}^{1}\!/_{2}\left(1 - \rho\left(A\right)\right)$ and apply theorem above.

Define $A^{*} = \overline{A^{T}}$ (complex conjugate transpose).

If A is Hermitian ($A = A^{*}$), then $\rho\left(A\right) = \left\|A\right\|_{2}$.

If A is normal ($AA^{*} = A^{*}A$),then $\rho\left(A\right) = \left\|A\right\|_{2}$.

Characteristic Polynomial

The eigenvalues of A are determined by the characteristic polynomial of A.

$$Ax = \lambda x \Leftrightarrow (A - \lambda I)x = 0$$

So we're looking for (eigen)values λ such that the matrix $(A - \lambda I)$ is singular:

$\det(A - \lambda I) = 0$ (this is a polynomial in λ)

This polynomial is called the characteristic polynomial of A. The eigenvalues of A are defined to be the roots of its characteristic polynomial.

Since eigenvalues of matrix are roots of its characteristic polynomial, the Fundamental Theorem of Algebra implies that an $n \times n$ matrix A always has n eigenvalues. The eigenvalues, however, need be neither distinct nor real.

Complex eigenvalues of a real matrix must come in complex conjugate pairs.

Multiplicity of eigenvalues

Eigenvalues may be single or multiple (single or multiple roots).
An eigenvalue with multiplicity $k > 1$ has k or fewer independent eigenvectors associated with it. It has at least one associated eigenvector. If it has fewer than k independent eigenvectors we call the eigenvalue (and the matrix) defective.

The multiplicity of an eigenvalue as the (multiple) root of the char. polynomial is called its algebraic multiplicity.
The number of independent eigenvectors associated with an eigenvalue is called its geometric multiplicity.

The geometric multiplicity is smaller than or equal to the algebraic multiplicity.

A matrix that is not defective is called diagonalizable: we have the decomposition

$$A = X\Lambda X^{-1} \Leftrightarrow X^{-1}AX = \Lambda = \mathrm{diag}(\lambda_i)$$

where X contains the eigenvectors (as columns) and Λ contains the eigenvalues.

Jordan form of a matrix

For every matrix $A \in \mathbb{C}^{n \times n}$ there exists a nonsingular matrix X such that

$$X^{-1}AX = \text{diag}(J_1, \cdots, J_q)$$

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix} \text{ and } J_i \in \mathbb{C}^{m_i \times m_i}, \text{ and } m_1 + m_2 + \cdots + m_q = n.$$

Each block has one corresponding eigenvector: q independent eigenvectors
Each block has $m_i - 1$ principal vectors (of grade 2)

If every block is of size 1, the matrix is diagonalizable
Multiple blocks can have the same eigenvalue: $\lambda_i = \lambda_j$
The sum of the sizes of all blocks with the same eigenvalue λ is the algebraic multiplicity of the eigenvalue λ. The number of blocks with the same eigenvalue λ is the geometric multiplicity of the eigenvalue λ.

Invariant Subspaces

A generalization of eigenvectors to higher dimensions is an invariant subspace.

We say a subspace $\mathcal{V} \subset \mathbb{C}^n$ is invariant under $A \in \mathbb{C}^{n\times n}$ if

for all $x \in V : Ax \in V$

It is possible that an invariant subspace is the span of a set of eigenvectors, but this need not be the case. Moreover, in general, elements (vectors) of the subspace will not be themselves eigenvectors.

In many cases it is useful to consider the restriction of the matrix A to the invariant subspace $\mathcal{V}$: $A_{\mathcal{V}} : \mathcal{V} \to \mathcal{V}$.

If $V \in \mathbb{C}^{n\times k}$ and $\textbf{range}(V) = \mathcal{V}$ then $AV = VL$ with $L \in \mathbb{C}^{k\times k}$.

Hence L represents A in the basis defined by V for the space $\mathcal{V}$

Similarity Transformation and Schur Decomposition

Let A have eigenpairs $\left(\lambda_i, v_i\right)$: $Av_i = \lambda_i v_i$

For nonsingular B, define the *similarity transformation*: BAB^{-1}

The matrix BAB^{-1} has the same eigenvalues, λ_i, as A and eigenvectors Bv_i:

$$BAB^{-1}\left(Bv_i\right) = BAv_i = \lambda_i Bv_i$$

In fact, BAB^{-1} has the same Jordan-block structure as A.

In many cases, we are interested in a (complex) unitary (or real, orthogonal) similarity transformation:

$$QAQ^* \qquad \text{with} \qquad QQ^* = Q^*Q = I$$

Schur decomposition: $QAQ^* = U$ (upper triangular)

Similarity Transformation

Similarity transformation for A: BAB^{-1};
this can be done with any nonsingular B.

Let $Ax = \lambda x$, then $BAB^{-1}Bx = BAx = \lambda Bx$.

BAB^{-1} has the same eigenvalues as A, and
eigenvectors Bx where x is an eigenvector of A.

Although any nonsingular B possible, most stable and
accurate algorithms with orthogonal (unitary) matrix.

For example used in the QR algorithm.

Similarity Transformation

Orthogonal similarity transformation for A: $Q^* A Q$, where $Q^* Q = I$.

If $Q^* A Q = \begin{pmatrix} L_1 & F \\ 0 & L_2 \end{pmatrix}$, then $\mathcal{L}(A) = \mathcal{L}(L_1) \cup \mathcal{L}(L_2)$.

If we can find $Q \equiv \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$ that yields such a decomposition we have reduced the problem to two smaller problems.

Moreover, $A Q_1 = Q_1 L_1$ and $\text{range}(Q_1)$ is invariant subspace.
Eigenpair $L_1 z = \lambda z$ gives eigenpair $A Q_1 z = Q_1 L_1 z = \lambda Q_1 z$.

Approximation over Search Space

For large matrices we cannot use full transformations. Often we do not need all eigenvalues/vectors. Look for proper basis Q_1 that captures relevant eigenpairs. We do not need Q_2.

Approximations over subspace $\mathrm{range}(Q_1)$: $L_1 = Q_1^* A Q_1$

When is an approximation good (enough)?

We will rarely find $A Q_1 - Q_1 L_1 = 0$ unless we do huge amount of work.
Not necessary. We are working with approximations and we must deal with numerical error anyway.

Approximation over Search Space

Let $AQ_1 - Q_1L_1 = R$ with $\|R\|$ small relative to $\|A\|$.

Now, $\left(A - RQ_1^*\right)Q_1 - Q_1L_1 = AQ_1 - R - Q_1L_1 = 0.$

$\text{range}(Q_1)$ is exact invariant subspace of perturbed matrix, $\hat{A}$.
$\hat{A} = A - RQ_1^*$ and $\left\|\hat{A} - A\right\| / \|A\| = \|R\| / \|A\|$

If $\|R\|/\|A\|$ sufficiently small, then Q_1 is acceptable.
In fact, this is as good as we can expect (unless we're lucky).
Any numerical operation involves perturbed operands!
Note that we cannot always say that Q_1 is accurate.

Eigenvalue problems

Before we compute eigenvalues and eigenvectors numerically, we must understand what we can and cannot compute (accurately) or should not compute.

We may want $Ax = \lambda x$ for single eigenpair, for example with λ as small as possible. Say minimum energy level.

In many cases we want $Ax_i = \lambda_i x_i$ for $i = 1 \ldots M$ (M large), where λ_i are smallest M eigenvalues.

- It may be important that we do not skip any eigenvalues.
- We may want the invariant subspace accurately.
- We may want every eigenvector accurately.

Usefulness of Computed Results

In general we need to consider the accuracy of a computed answer, without knowing the exact answer.

This involves the sensitivity of the result we want to compute.

If some result is very sensitive to small changes in the problem, it may be impossible to compute exactly. In other cases results may be computable but at very high price, for example, an algorithm may convergence very slowly.

Sometimes it is better to compute related but less sensitive result.

Sensitivity of an Eigenvalue

Sensitivity of eigenvalues to perturbations in the matrix.

Different eigenvalues or eigenvectors of a matrix are not equally sensitive to perturbations of the matrix.

Let $Ax = \lambda x$ and $y^* A = \lambda y^*$, where $\|x\| = \|y\| = 1$.

Consider $(A+E)(x+e) = (\lambda + \varepsilon)(x+e)$ and drop second order terms.

$$Ax + Ae + Ex \cong \lambda x + \lambda e + \varepsilon x \Leftrightarrow Ae + Ex \cong \lambda e + \varepsilon x$$

$$y^* Ae + y^* Ex = \lambda y^* e + y^* Ex \cong \lambda y^* e + \varepsilon y^* x \quad \Leftrightarrow$$

$$y^* Ex \cong \varepsilon y^* x \Rightarrow \varepsilon \cong \frac{y^* Ex}{y^* x} \Rightarrow |\varepsilon| \leq \frac{\|E\|}{y^* x}$$

Condition number of simple eigenvalue: $\kappa(\lambda) = \left| y^* x \right|^{-1}$

Sensitivity of an Eigenvalue

For symmetric/Hermitian matrix, right and left eigenvectors are the same. So, eigenvalues are inherently well-conditioned.

More generally, eigenvalues are well conditioned for normal matrices, but eigenvalues of nonnormal matrices need not be well conditioned.

Nonnormal matrices may not have a full set of eigenvectors. The algebraic multiplicity, the multiplicity of λ as a root of $\det(A-\lambda I)=0$, is not equal to the geometric multiplicity, $\dim \operatorname{null}(A-\lambda I)$. In that case we can consider conditioning of the invariant subspace associated with a Jordan block.

Sensitivity of an Eigenvalue

If μ is an eigenvalue of $A+E$, then $\lambda \in \mathcal{L}(A)$ exists:

$$|\mu-\lambda| \le \left\|XEX^{-1}\right\| \le \kappa(X)\|E\|$$

where X is the matrix of eigenvectors of A and $\kappa(X) \equiv \|X\|\left\|X^{-1}\right\|$ is a condition number (consistent norm).

A useful backward error result is given by the residual.

Let $r = Ax - \lambda x$ and $\|x\| = 1$.

Then there exists a perturbation E with $\|E\| \le \|r\|$ such that

$$(A+E)x = \lambda x.$$

Proof: Take $E = -rx^{*}$.

Sensitivity of Eigenvectors

Consider $A = \begin{pmatrix} 1 & 0 \\ 0 & 1+\varepsilon_1 \end{pmatrix} \quad \rightarrow \quad x_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ x_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Consider perturbation E with $\|E\| = \varepsilon_1 + \varepsilon_2$, ε_2 arb. small.

Enough to give A any eigenvectors not equal to x_1 and x_2.

Let $E = E_1 + E_2$ and $\hat{X} = \begin{bmatrix} \hat{x}_1 & \hat{x}_2 \end{bmatrix}$ (unitary)

Let $E_1 = \begin{pmatrix} 0 & 0 \\ 0 & -\varepsilon_1 \end{pmatrix}$ and $E_2 = \hat{X} \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon_2 \end{pmatrix} \hat{X}^*$.

$A + E_1 = I$ (all nonzero vectors are eigenvectors)

$$A + E = I + \hat{X} \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon_2 \end{pmatrix} \hat{X}^* = \hat{X} \begin{pmatrix} 1 & 0 \\ 0 & 1+\varepsilon_2 \end{pmatrix} \hat{X}^*$$

SVD

$A^{n\times m} = U \Sigma V^*$ $\quad U^{n\times n}$ unitary

$V^{m\times m}$ unitary

$\Sigma^{n\times m}$ diag, real, nonneg.

$= U \begin{pmatrix} \sigma_1 & & \\ & \ddots & \sigma_m \\ & 0 & \end{pmatrix} V^*$ or $U \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_m & \end{pmatrix} V^*$

or $U \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \end{pmatrix} V^*$ (square)

$\sigma_i \in \mathbb{R} \geq 0$

convention : $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m$

sometimes $\sigma_1 \geq \cdots \geq \sigma_p$ $\quad p = \min(m, n)$

or better $p = \operatorname{rank}(A)$

$A = U^{n\times m} \Sigma^{m\times m} (V^{m\times m})^*$ $\quad$ economy SVD $(\cancel{n}\ m \geq n)$

or $U^{n\times n} \Sigma^{n\times n} \overset{n\times m}{V^*}$

If A real U, V real orthog.

$R(U_p) = R(A)$ $\quad A = U^{n\times p} \Sigma^{p\times p} (V^*)^{p\times m}$

$R(A) = u_1 \cdots u_p$ $\quad \operatorname{null}(A) = \cancel{R}\ \operatorname{span}\{v_{p+1}, v_m\}$

$R(A)^\perp = u_{p+1} \cdots u_n$ $\quad \operatorname{null}(A)^\perp = v_1 \cdots v_p$

$A = \sum \sigma_i u_i v_i^*$ $\quad A^+ = \sum_{\substack{\cancel{\sigma \neq}\, i: \\ \sigma_i > 0}} \sigma_i^{-1} v_i u_i^* = V \Sigma^+ U^*$

$$x = A^{-1}b = \sum_{i=1}^{n} v_i \frac{u_i^* b}{\sigma_i}$$

(note sensitivity in u_i directions corr. to smallest σ_i)

$$\|A\|_2 \, \|A^{-1}\|_2 = \sigma_1/\sigma_n \quad \text{or} \quad \sigma_{max}/\sigma_{min}$$

$$\|A\|_2 = \sigma_1 \quad (\max \sigma_i)$$

$$\|A^{-1}\|_2 = \max\left(\frac{1}{\sigma_i}\right) = \sigma_{min}^{-1} = \sigma_n^{-1}$$

$$\|A\|_F = \left(\sum_i \sigma_i^2\right)^{1/2}$$

$$\min_{A+E \text{ sing}} \|E\|_2 = \sigma_{min}$$ → distance to singularity

---

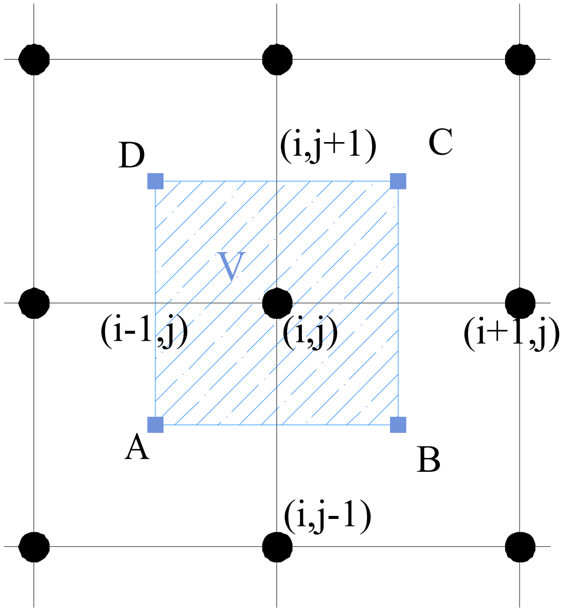
with some effort (see notes corr. Watkins)

$$\frac{\|\Delta x\|_2}{\|x\|} \leq \left(\text{cond}(A)_2^2 \tan\theta + \text{cond}_2(A)\right) \frac{\|E\|_2}{\|A\|_2}$$

where θ $\measuredangle\left(b, \text{range}(A)\right)$

Model Problems

Discretize $-\left(pu_x\right)_x - \left(qu_y\right)_y + ru_x + su_y + tu = f$.



Integrate equality over box V. Use Gauss' divergence theorem to get

$$\int_V \left(pu_x\right)_x + \left(qu_y\right)_y \, dxdy = \int_{\partial V} \begin{pmatrix} pu_x \\ qu_y \end{pmatrix} \cdot n \, ds$$

And approximate the line integral numerically.

Model Problems

Now we approximate the boundary integral $\int_{\partial V} \begin{pmatrix} pu_x \\ qu_y \end{pmatrix} \cdot n\, ds$.

We approximate the integrals over each side of box V using the midpoint rule and we approximate the derivatives using central differences.

$$\int_B^C pu_x n_1 dy \approx \frac{\Delta y}{\Delta x} p_{i+1/2,j} \left(U_{i+1,j} - U_{i,j}\right)$$ and so on for the other sides

We approximate the integrals over ru_x, su_y, tu, and f using the area of the box and the value at the midpoint of the box, where we use central differences for derivatives. So, $u_x \approx \left(U_{i+1,j} - U_{i-1,j}\right) / \left(2\Delta x\right)$, and so on.

For various examples we will also do this while strong convection relative to the mesh size makes central differences a poor choice (as it gives interesting systems).

Model problems

This gives the discrete equations

$$
\begin{aligned}
&-\frac{\Delta y}{\Delta x}\left[p_{i+1/2,j}\left(U_{i+1,j}-U_{i,j}\right)-p_{i-1/2,j}\left(U_{i,j}-U_{i-1,j}\right)\right]\\
&-\frac{\Delta x}{\Delta y}\left[q_{i,j+1/2}\left(U_{i,j+1}-U_{i,j}\right)-p_{i,j-1/2}\left(U_{i,j}-U_{i,j-1}\right)\right]\\
&+\left(\Delta y / 2\right)r_{i,j}\left(U_{i+1,j}-U_{i-1,j}\right)+\left(\Delta x / 2\right)s_{i,j}\left(U_{i,j+1}-U_{i,j-1}\right)\\
&+\Delta x\Delta y t_{i,j}U_{i,j}=\Delta x\Delta y f_{i,j}
\end{aligned}
$$

Often we divide this result again by $\Delta x\Delta y$.

Rate of Convergence

Let $\hat{x}$ be the solution of $Ax = b$, and we have iterates $x_0, x_1, x_2, \dots$

$\{x_k\}$ converges (q-)linearly to $\hat{x}$ if there are $N \geq 0$ and $c \in [0,1)$ such that for $k \geq N: \| x_{k+1} - \hat{x} \| \leq c \| x_k - \hat{x} \|$,

$\{x_k\}$ converges (q-)superlinearly to $\hat{x}$ if there are $N \geq 0$ and a sequence $\{c_k\}$ that converges to 0 such that for $k \geq N: \| x_{k+1} - \hat{x} \| \leq c_k \| x_k - \hat{x} \|$

$\{x_k\}$ converges to $\hat{x}$ with (q-)order at least p if there are $p > 1$, $c \geq 0$, and $N \geq 0$ such that $k \geq N: \| x_{k+1} - \hat{x} \| \leq c \| x_k - \hat{x} \|^p$ (quadratic if $p = 2$, cubic if $p = 3$, and so on)

$\{x_k\}$ converges to $\hat{x}$ with j-step (q-)order at least p if there are a fixed integer $j \geq 1$, $p > 1$, $c \geq 0$, and $N \geq 0$, such that $k \geq N: \| x_{k+j} - \hat{x} \| \leq c \| x_k - \hat{x} \|^p$