Apuntes de clases 2020B

Análisis Numérico Lineal

Evaluación:

* Teoría (/10) (70%) — 1 Prueba /5; 1 Examen /5
* Laboratorio (/10) (30%) — Laboratorios /5; Expo / Trabajo /5

Ideas:

$\boxed{Ax=b}$!!!

$A \in \mathbb{R}^{m\times n}$, $A \in \mathbb{R}^{n\times n}$

$x \in \mathbb{R}^n$

$b = \mathbb{R}^m$

→ Anális de Errores

↳ Condicionamiento del problema

↳ Estabilidad del Algoritmo

→ Métodos ($Ax=b$)

↳ Directos (Trabajo directo sobre los elementos de la matriz A.

↳ A (Descomposiciones de A)

↳ Ej: $A = \underset{\triangle}{L}\,\underset{\triangledown}{U} \Rightarrow LUxb$

↳ Iterativos

$(x^0, x^1, x^2, \ldots\ldots) \longrightarrow \bar{x}$

→ Valores y Vectorios Propios (A).

Bibliografía:

* A. Quarteroni, Numerical Mathematics. Springer
* A. " ". S. Saleri, Scientific Computing with Matlab and octave.

E-mail: sofia.lopezo@epn.edu.ec

Tutoría: Jueves 6-8 pm

Nociones de normas

Def: Un espacio normado es un espacio V sobre $\mathbb{R}$, equipado con $\|\cdot\|_V : V \longrightarrow \mathbb{R}^+ \cup \{0\}$ llamada norma y satisface:

✓ a) $\|v\|_V \geq 0 \quad \forall v \in V$

b) $\|v\| = 0 \iff v = 0$

✓ c) $\|\alpha v\|_V = |\alpha|\,\|v\|_V, \ \forall v \in V$ y $\forall \alpha \in \mathbb{R}$

✓ d) $\|v+w\|_V \leq \|v\|_V + \|w\|_V \ \forall v, w \in V$.

Nota: Si sólo se cumple a), c) y d) $\|\cdot\|$ es una semi-norma.

Def 2. Sea V un espacio, donde se han definido dos normas: $\|\cdot\|_p$ y $\|\cdot\|_q$. Se dice que éstas son equivalentes si existen dos constantes α_{pq} y $\beta_{pq} \in \mathbb{R}^+$, tq:

$$\alpha_{pq}\|x\|_q \leq \|x\|_p \leq \beta_{pq}\|x\|_q \qquad \forall x \in V$$

equivalentes si existen dos constantes $\alpha_{pq}, \beta_{pq} \in \mathbb{R}$, tq.

$$\alpha_{pq}\|x\|_q \leq \|x\|_p \leq \beta_{pq}\|x\|_q \quad \forall x \in V$$

Def 3. Un operador lineal es una aplicación A, tq. $A: (V, \|\cdot\|_V) \to (X, \|\cdot\|_X)$ que cumple la propiedad:

$$A(\alpha v + \beta w) = \alpha Av + \beta Aw, \quad \forall v, w \in V \text{ y } \forall \alpha, \beta \in \mathbb{R}$$

Nota: En dimensión finita, una aplicación lineal $A: \mathbb{R}^n \to \mathbb{R}^m$ siempre está representada por una única matriz $A \in \mathbb{R}^{m \times n}$ y viceversa.

Def 4. A es un operador acotado si existe una constante $K > 0$, independiente de v tq.

$$\|Av\|_X \leq K\|v\|_V \quad \forall v \in V$$

Def 5: Sea V un espacio. Una aplicación $a: V \times V \to \mathbb{R}$ es llamada forma bilineal si cumple las sig. propiedades:

a) $a(\alpha v + \beta w, x) = \alpha a(v,x) + \beta a(w,x) \quad \forall (v,w,x) \in V \times V \times V$ y $\alpha, \beta \in \mathbb{R} \times \mathbb{R}$

b) $a(v, \alpha x + \beta y) = \alpha a(v,x) + \beta a(v,y) \quad \forall (v,x,y) \in \|\cdot\|$ y

si a es una forma bilineal en $(V, \|\cdot\|)$ entonces a es:

* acotada: si existe una constante $K > 0$, (ind. de v, w) tq.

$$|a(v,w)| \leq K\|v\|\|w\|, \quad \forall (v,w) \in \|\cdot\|$$

* simétrica: $a(v,w) = a(w,v)$, $\|\cdot\|$
* antisimétrica: $a(v,w) = -a(w,v)$, $\|\cdot\|$
* semidefinida positiva: si $a(v,v) \geq 0$, $\forall v \in V$
* definida positiva: si $a(v,v) > 0$ $\forall v \in V$ y $v \neq 0$
* coerciva: si existe una constante $c > 0$ tq. $a(v,v) > c\|v\|^2$, $\forall v \in V$.
* producto escalar: si a es simétrica y definida positiva y en este caso se escribe $\langle \cdot, \cdot \rangle$ en lugar de $a(\cdot,\cdot)$.

La aplicación $: X \mapsto \sqrt{\langle x, x \rangle}$ definida en un espacio X con un producto escalar, define una norma en X, llamada norma inducida.

Desigualdad de Cauchy - Schwarz:

$$|\langle v, w \rangle| \leq \|v\|\|w\| \quad \forall v, w \in V$$

Ej. *Consideremos la forma bilineal a definida en $\mathbb{R}^n$ como:

$$a(x,y) = x^T y = \sum_{i=1}^{n} x_i y_i$$

la cual coincide con el producto escalar

usual de $\mathbb{R}^n$

* Si $B \in \mathbb{R}^{n\times n}$, se define la forma bilineal
$a(x,y) = x^T B y > 0$

(Nota: Si B es s.d.p, entonces se puede definir un producto escalar en $\mathbb{R}^n$)

Normas vectoriales:

En $\mathbb{R}^n$ se definen las siguientes normas:

* Normas Hölderianas:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad 1 \le p < \infty$$

Si $p=2$ tenemos la norma Euclideana en $\mathbb{R}^n$.

* Norma uniforme: $\|x\|_\infty = \max_{1\le i\le n} |x_i|$

Desigualdad de Young: Sean dos números reales no negativos a y b entonces:

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}$$

para cualesquiera p y q no negativos tq $\frac{1}{p} + \frac{1}{q} = 1$

Desigualdad de Hölder:

$$|\langle x, y\rangle| \le \|x\|_p \|y\|_q \quad \forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n$$

con p y q tq: $\frac{1}{p} + \frac{1}{q} = 1$.

Desigualdad de Minkowski:

$$\|x+y\|_p \le \|x\|_p + \|y\|_p \quad \forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n$$
$$\text{y } p \in [1,\infty]$$

Equivalencia de normas: Las constantes de equivalencia α_{pq} y β_{pq} (ver def 2) para distintas normas vectoriales definidas en $\mathbb{R}^n$ se encuentran en la siguiente tabla:

α_{pq}	$q=1$	$q=2$	$q=\infty$
$p=1$	1	1	1
$p=2$	$\frac{1}{\sqrt{n}}$	1	1
$p=\infty$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$	1

β_{pq}	$q=1$	$q=2$	$q=\infty$
$p=1$	1	$\sqrt{n}$	n
$p=2$	1	1	$\sqrt{n}$
$p=\infty$	1	1	1

Normas matriciales:

En $\mathbb{R}^{n\times n}$ se definen las siguientes normas:

* Normas Hölderianas: (inducidas)

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

$$= \sup_{\|x\|\le 1} \|Ax\|_p$$

Nota: En general $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$

* Norma de Frobenius

$$\|A\|_F = \sqrt{\sum_{i,j}^{n} |a_{ij}|^2} = \sqrt{tr(A^T A)} = \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2}$$

Donde σ_i (valores singulares de A = raíz cuadrada de los vp. de $A^T A$)

* Norma Espectral: $\|A\| = \sqrt{\rho(A^T A)}$

($\rho(B) = \max_i |\lambda_i|$, λ_i vp de B)

* $\|A\|_\infty = \max_{i=i,\dots,m} \sum_{j=1}^{n} |a_{ij}|$ (inducida por la norma vec. $\|\cdot\|_\infty$)

↳ máximos de la suma de las filas.

* $\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$ (inducida por la norma vec $\|\cdot\|_1$)

↳ máximo de la suma de las columnas.

Propiedades: Sea $\|\cdot\|_M$ una norma matricial y $\|\cdot\|_{V_1}$, $\|\cdot\|_{V_2}$ dos normas vectoriales, decimos que $\|\cdot\|_M$ es compatible o consistente con las normas vectoriales $\|\cdot\|_{V_1}$ y $\|\cdot\|_{V_2}$ si

$$\|Ax\|_{V_1} \leq \|A\|_M \|x\|_{V_2}.$$

Por otro lado, decimos que $\|\cdot\|_M$ es submultiplicativa

$$\|AB\|_M \leq \|A\|_M \|B\|_M$$

¿Qué normas cumplen estas propiedades?

↳ Las normas matriciales inducidas por una norma vectorial son siempre compatibles

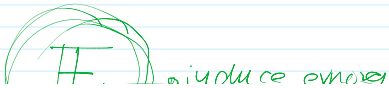
↳ Si $\|\cdot\|$ es una norma matricial inducida, entonces $\|I\| = 1$.

↳ Si $\|\cdot\|$ es una norma matricial inducida, entonces $\|AB\| \leq \|A\|\|B\|$.

Constantes de equivalencia

α_{pq}	$q=1$	$q=2$	$q=\infty$
$p=1$	1	$\frac{1}{\sqrt{n}}$	$\frac{1}{n}$
$p=2$	$\sqrt{m}$	1	$\frac{1}{\sqrt{n}}$
$p=\infty$	$\frac{1}{m}$	$\frac{1}{\sqrt{m}}$	1

β_{pq}	$q=1$	$q=2$	$q=\infty$
$r=1$	1	$\sqrt{n}$	n
$r=2$	$\sqrt{n}$	1	$\sqrt{m}$
$p=\infty$	n	$\sqrt{n}$	1



$p=\infty$

$\mathbb{F}$ → induce errores

Análisis de errores:

↳ condicionamiento problema!

↳ Estabilidad (algoritmos)!

Ej: Supongamos que tenemos una calculadora con 4 dígitos decimales y queremos resolver el sig. sistema:

$$\begin{cases} \frac{1}{10000} x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

$x_1 = 1{,}0001$

$x_2 = 0{,}9999$

$1e{-4}\, x_1 + 1{,}0000\, x_2 = 1{,}0000$

$1{,}0000\, x_1 + 1{,}0000\, x_2 = 2{,}0000$

Multiplicamos la 1era ec. por $-10\,000$ y sumamos a la segunda

$$\begin{array}{rl} -1{,}0000\, x_1 - 1e{+4}\, x_2 &= -1e{+4} \\ +\quad 1{,}0000\, x_1 + 1{,}0000\, x_2 &= 2{,}0000 \\ \hline -9999\, x_2 &= -9998 \end{array}$$

$$\boxed{\tilde{x}_2 = 0{,}9998}$$

Reemplazando en la 1era ec. tenemos:

$$\frac{1}{10000} x_1 + 1{,}0000 \cdot 0{,}9998 = 1{,}0000$$

$$\boxed{\tilde{x}_1 = 2}$$

No satisface el sistema.

Sin embargo, la sol exacta del sistema es:

$$\boxed{\begin{array}{l} x_1 = 1{,}0000 \\ x_2 = 0{,}9999 \end{array}}$$

Veremos que al resolver un problema numérico los errores cometidos pueden ser analizados estudiando la naturaleza del problema y por otro lado, el método de solución. Así, podemos analizar las distintas fuentes de errores:

DATOS → ALGORITMO = Resultado

- errores de redondeo
- errores de truncamiento
- mediciones imprecisas

errores pequeños + errores pequeños = errores pequeños?

Definición: Sea $\bar{x} \in \mathbb{R}$ una sol. exacta y $\hat{x}$ un valor aprox. a $\bar{x}$, entonces definimos:

Error absoluto: $|\bar{x} - \hat{x}|$

→ Error relativo: $\frac{|\bar{x} - \hat{x}|}{\ }$, $\bar{x} \neq 0$

Definición $\bar{x}$, aproximación $\hat{x}$

Error absoluto: $|\bar{x} - \hat{x}|$

→ Error relativo: $\frac{|\bar{x} - \hat{x}|}{|\bar{x}|}$, $\bar{x} \neq 0$

Diferencia entre el error absoluto y relativo

① $\bar{x} = 0.5 \times 10^{-4}$

$\hat{x} = 0.4 \times 10^{-4}$

error absoluto $|\bar{x} - \hat{x}| = 0.1 \times 10^{-4}$

error relativo $\frac{|\bar{x} - \hat{x}|}{|\bar{x}|} = 0.2$ (20%)

② $\bar{x} = 5000$

$\hat{x} = 4950$

error abs $|\bar{x} - \hat{x}| = 50$

error relat. $\frac{|\bar{x} - \hat{x}|}{|\bar{x}|} = 0.01$ (1%)

↳ Se interpreta como un error en porcentaje.

Números de punto flotante

Como un computador tiene recursos limitados, entonces solo un subconjunto $\mathbb{F}$ de dimensión finita de $\mathbb{R}$ puede ser representado en una máquina. Estos números ($\mathbb{F}$) se denominan números de punto flotante. Más aún, las propiedades de $\mathbb{F}$ son diferentes a las de $\mathbb{R}$, la razón es que cualquier número real x es, en principio, truncado por la máquina, dando como resultado un nuevo número, denotado por $fl(x)$, que no necesariamente coincide con el original. (-)

Ej: Consideremos $x = \frac{1}{7} = 0.\overline{142857}$. La computadora los va a regresar.

0.1429

Este mismo comportamiento se da si tuvimos $1/3$,

0.3333.

Es decir, con 4 dígitos decimales, la aproximación no es exacta, puesto que, la computadora está redondeando los números. Esto significa que sólo una cantidad fija de dígitos decimales puede ser retornada. La representación interna del número se realiza con hasta 16 dígitos decimales.

En general, una computadora almacena un número real de la siguiente manera:

$$x = (-1)^s \cdot (0.a_1 a_2 a_3 \ldots a_t) \cdot \beta^e$$
$$= (-1)^s \cdot m \cdot \beta^{e-t}, \quad |a_1 \neq 0|,$$

s es 0 o 1, β es un número entero positivo mayor o igual a 2 a la base, m es un entero

Sea $0.d_1 d_2 \dots d_t$, β (es un número entero positivo mayor o igual 2) es la base, m es un entero llamado mantissa cuya longitud t es el máximo número de dígitos a_i que puede almacenarse ($0 \le a_i \le \beta - 1$). Por último e es un entero llamado el exponente.

Como la posición del punto decimal no es fija, se denominan números de punto flotante. Los dígitos $a_1 a_2 \dots a_p$ ($p \le t$) se denominan los p primeros dígitos significativos de x.

La condición $a_1 \neq 0$ asegura que el número no tenga múltiples representaciones:

Ej: $\frac{1}{10}$

(base decimal) $\cdot\ 0.1 \cdot 10^{0} = 1/10$

$0.01 \cdot 10^{1} = 1/10$ $\Rightarrow$ es por esto $a_1 \neq 0$

$0.001 \cdot 10^{2} = 1/10$

El conjunto $\mathbb{F}$ se caracteriza por la base β, el número de dígitos significativos t y el rango (L, U) ($L < 0$ y $U > 0$) de variación del índice e.

Así se denota $\mathbb{F}(\beta, t, L, U)$.

Ej: En Matlab $\mathbb{F}(2, 53, -1021, 1024)$.

En ese caso, 53 dígitos significativos en base 2 corresponden a 15 dígitos significativos en base 10 que son mostrados en Matlab (usando format long).

Errores de redondeo: Como es imposible no generar errores al reemplazar x por $fl(x)$, tenemos la siguiente estimación para el error:

$$\frac{|x - fl(x)|}{|x|} \le \frac{1}{2} \varepsilon_M$$

donde $\varepsilon_M = \beta^{1-t}$, se denomina epsilon de la máquina y mide la distancia entre 1 y el número de punto flotante más cercano y mayor a 1.

Ej: En Matlab, $\varepsilon_M = \beta^{1-t} = 2^{-52}$

El número 0 no pertenece a $\mathbb{F}$, en este caso deberíamos tener $a_1 = 0$. Más aún, como L y U son finitos, no podemos representar números cuyo valor absoluto sea "demasiado grande" o "demasiado pequeño". Precisamente, el menor y mayor números reales de $\mathbb{F}$ están dados por:

$\rightarrow x_{min} = \beta^{L-1}$

reales de $\mathbb{F}$ están dados por:

→ $x_{min} = \beta^{L-1}$

→ $x_{max} = \beta^{U}(1-\beta^{-t})$

Un número positivo más pequeño x_{min} produce un mensaje de underflow y se lo trata como 0.

Un número positivo mayor a x_{max} produce un error de overflow y se almacena en la variable Inf ($+\infty$).

Operaciones de punto flotante

Existen ciertas propiedades en $\mathbb{R}$ que no se satisfacen en $\mathbb{F}$. Por ejemplo la propiedad conmutativa de la suma sí se satisface: $fl(x+y) = fl(y+x)$.

De igual manera, se satisface para la multiplicación: $fl(x \cdot y) = fl(y \cdot x)$.

Sin embargo la propiedad distributiva y asociativa NO se satisfacen.

Existe un error en particular que se da cuando se suman 2 números con signos opuestos pero valor absoluto similar, en este caso, el resultado es incorrecto y se refiere a un error de pérdida o cancelación de dígitos significativos

Ej: $((1+x)-1)/x$

$x = 1.e-15$

ans = 1.1102

→ Ariane 5 (1996).

Resumen: $\mathbb{F} \subset \mathbb{R}$

- Finito
  - x_{min} (error más pequeño que se puede representar en número de punto fl.)
  - x_{max} (" " grande)
  - $0 \notin \mathbb{F}$
  - $x = (-1)^s \cdot m \cdot \beta^{e-t}$ (truncado)

- errores
  - overflow
  - underflow
  - No se satisfacen todas las prop. de $\mathbb{R}$. Ej: asoc. y dist.

Operaciones de punto flotante

Consideremos las operaciones aritméticas $\{+, -, \cdot, \div\}$ y sus versiones respectivas de p.f. $\{\hat{+}, \hat{-}, \hat{\cdot}, \hat{\div}\}$.

En general, supondremos que el error relativo que se comete al realizar una operación aritmética de punto flotante $\circ \in \{+,-,\cdot,\div\}$ entre dos números x e y en $\mathbb{F}$ es de $\varepsilon = \varepsilon(x,y)$ y satisface el siguiente modelo:

$$\left| \frac{x \hat{\circ} y - x \circ y}{x \circ y} \right| = |\varepsilon| \leq eps$$

números x y y en $\mathbb{F}$ es de ε [illegible] ε

satisface el sig. modelo:

$$\rightarrow \boxed{\frac{x \hat{o} y - x o y}{x o y} = \varepsilon} \quad \text{con } |\varepsilon| \leq eps$$

A cumulación de errores de redondeo

- Si dos números x y y en $\mathbb{F}$ difieren significativamente en magnitud $x+y$, tiene un error absoluto grande.
- Si $|y| << 1$, entonces x/y puede provocar errores abs. (o relativos) muy grandes.
- Si $|x| \approx |y|$ y tienen el mismo signo $x-y$ tiene un error relativo muy grande, conocido como error de cancelación.

Análisis de condición normado:

Notación o y O.

Sean f y g dos funciones definidas en un abierto U subconjunto de un espacio normado V en otro espacio normado X. Utilizaremos la sig. notación:

- $f(x) = O(g(x))$ cuando $x \to a$, si se verifica que $\|f(x)\| = C\|g(x)\|$ para cierta constante C y $|x-a| < \delta$.
- $f(x) = o(g(x))$ cuando $x \to a$, significa que $\lim_{x \to a} \frac{\|f(x)\|}{\|g(x)\|} = 0$

Ej: Consideremos la expansión de Taylor de orden 2 de la fun. sen. alrededor de 0.

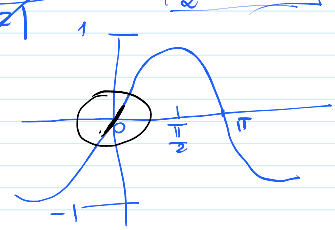
$$\operatorname{sen}(x) = \operatorname{sen}(0) + \cos(0)(x) - \underbrace{\frac{1}{2}\operatorname{sen}(\xi)(x-0)^2}_{o(|x|),\ O(|x|^2)} \quad \text{con } \xi \in [0,x]$$

$$\Rightarrow \frac{1}{2}\operatorname{sen}(\xi)(x)^2 = o(|x|)$$

i.e, $$\lim_{x \to 0} \frac{\frac{1}{2}|\operatorname{sen}(\xi)|\,|\cancel{x^2}|}{|\cancel{x}|} = \lim_{x \to 0} \frac{1}{2}|\operatorname{sen}(\xi)|\,|x| = 0$$

$$\Rightarrow \frac{1}{2}\operatorname{sen}(\xi)(x^2) = O(|x|^2)$$

$$\frac{\frac{1}{2}|\operatorname{sen}(\xi)|\,|\cancel{x^2}|}{|\cancel{x^2}|} = \overset{C}{\boxed{\frac{1}{2}|\operatorname{sen}(\xi)|}}$$



Así, concluimos que $\operatorname{sen}(x) \to x$ cuando $x \to 0$, es decir, la función seno se comporta como la

Así, concluimos que $\text{sen}(x) \longrightarrow x$ cuando $x \to 0$, es decir, la función seno se comporta como la función x alrededor de 0, excepto por el residuo $\frac{1}{2}\text{sen}(\xi)x^2$ que se considera como el error de orden cuadrático al aproximar $\text{sen}(x)$ por la función x (identidad) cerca de 0.

Utilizar la notación

$$\text{sen}(x) \doteq x \iff \text{sen}(x) = x + O(|x|)$$

Def Dadas dos funciones $g, h : \mathbb{R}^n \to \mathbb{R}^m$, se dice que son iguales en orden para $x \to x_0$ ssi, $g(x) = h(x) + O(\|h(x)\|)$ y se denota

$$g(x) \doteq h(x) \quad \text{para } x \to x_0$$

Def Análogamente a la igualdad en orden, para g y $h : \mathbb{R}^n \to \mathbb{R}^m$ tenemos:

$$g(x) \dot{\leq} h(x) \quad \text{para } x \to x_0$$

$$\iff g(x) \leq h(x) + O(\|h(x)\|)$$

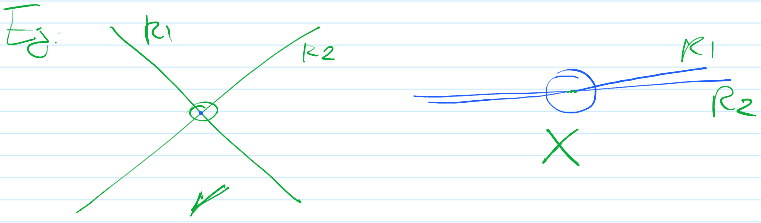
Ejemplo: Si f es diferenciable en x

$$f(\hat{x}) \doteq f(x) - f'(x)(\hat{x} - x)$$

cuando $x \to \hat{x}$.

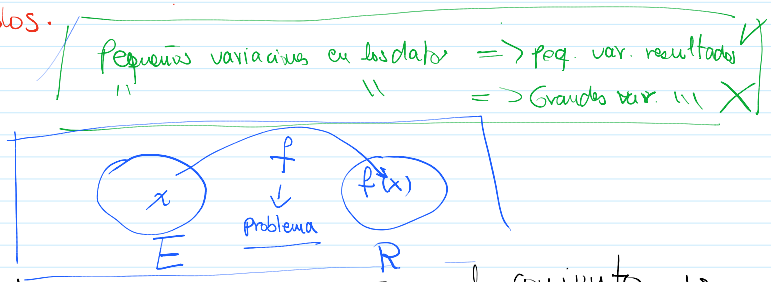
Condicionamiento de un problema:
Se refiere al efecto que cambios en los datos reflejan en los resultados, independientemente de la forma en que los últimos son obtenidos.

Ej:



En el ejemplo anterior, mientras uno de los ángulos entre ambas rectas es cada vez más pequeño, las variaciones que pueden ocurrir en los datos tienen un efecto más visible en los resultados. En este caso, es el ángulo el que caracteriza este comportamien-

visible en los resultados. En este caso, es el algoritmo el que caracteriza este comportamiento. La condición de un problema, justamente, corresponde a un número que nos indica cómo la variación de los datos influye en los resultados.



Consideremos, para $x \in \mathbb{F}$, el conjunto de datos con precisión relativa:

$$E = \left\{ \tilde{x} \in \mathbb{R}^u : \frac{\|\tilde{x} - x\|}{\|x\|} \leq eps \right\}$$

donde eps corresponde al epsilon de la máquina.

Y, por otro lado, tenemos el conjunto con precisión absoluta:

$$E_a = \{ \tilde{x} \in \mathbb{R}^u : \|\tilde{x} - x\| \leq \delta x \}$$

con $\delta x > 0$.

Sea $f : E \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ una función que asocia a cada x en E la solución de un problema que se encuentra en $R = f(E)$, entonces la condición del problema (f, x) es una medida que caracteriza la relación entre E y R.

Def La condición absoluta en norma del problema (f, x) es la constante más pequeña K_{abs}, tq.

$$\|f(\hat{x}) - f(x)\| \leq K_{abs} \|\hat{x} - x\|.$$

para $\hat{x} \to x$ — Número de condición

De la misma manera, la condición relativa en norma es la constante más pequeña, K_{rel}, tq.

$$\frac{\|f(\hat{x}) - f(x)\|}{\|f(x)\|} \leq K_{rel} \frac{\|\hat{x} - x\|}{\|x\|}$$

para $\hat{x} \to x$

Obs: • Si no existe K_{abs} finito, entonces el problema está mal definido

• K_{abs} y K_{rel} son una medida de proporcionalidad de los errores en las soluciones con respecto a los errores en los datos

• Un problema es bien condicionado cuando su número de condición absoluto o relativo es "pequeño". La noción de pequeño está determinada por las aplicaciones.

Resultado (★): Si f es F-diferenciable en x se puede caracterizar K_{abs} de la sig manera:

$$K_{abs} = \|f'(x)\|$$

$$f(x) = f(\hat{x}) - f'(x)(x-\hat{x}) + O(\|x-\hat{x}\|)$$

$$\|f(x) - f(\hat{x})\| \leq \|-f'(x)(x-\hat{x})\| + \|O(x-\hat{x})\|$$

$$\leq \underbrace{\|f'(x)\|}_{K_{abs}} \|x-\hat{x}\|,$$

donde, $\|f'(x)\|$ es la norma de la matriz jacobiana de f, usando la norma matricial, $\|f'(x)\| = \sup_{y \neq 0} \frac{\|f'(x)\,y\|}{\|y\|}$.

El K_{rel}, se caracteriza de la sig. manera.

$$K_{rel} = \frac{\|f'(x)\|\,\|x\|}{\|f(x)\|} =$$

$$= \frac{K_{abs}\,\|x\|}{\|f(x)\|}$$

Ejemplo: Calcular el número de condición de la suma y resta dos números reales.

$$f: \mathbb{R}^{(2)} \to \mathbb{R}^{(1)}$$
$$(a,b) \mapsto x = a+b.$$

Debemos calcular $\|f'(a,b)\|$.

Así: $f'(a,b) = \nabla f(a,b)^T = \begin{pmatrix} 1 \\ 1 \end{pmatrix}^T$.

(f es diferenciable)

Así: $f'(a,b) = \underset{\substack{\uparrow \\ f \text{ es diferenciable}}}{\mathbb{1}} f(a,b) \quad \begin{pmatrix} 1 \end{pmatrix}$

$$K_{abs_+} = \|f'(a,b)\| = \sup_{(x_1,x_2)\neq(0,0)} \frac{\left\| f'(a,b)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right\|}{\left\|\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right\|}$$

$$= \sup_{x\neq 0} \frac{\left\|\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right\|_1}{\left\|\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right\|_1}$$

$$= \sup_{x\neq 0} \frac{|x_1+x_2|}{|x_1|+|x_2|} = 1$$

Por otro lado, $\dfrac{|x_1+x_2|}{|x_1|+|x_2|} \leq \dfrac{|x_1|+\cancel{|x_2|}}{|x_1|+\cancel{|x_2|}} \quad 1$

Así $K_{abs_+} = 1$

$$K_{rel_+} = \frac{1 \cdot \|(a,b)\|_1}{|a+b|} = \frac{|a|+|b|}{|a+b|}$$

Para la resta:

$K_{abs_-} = 1$

$$K_{rel_-} = \frac{|a|+|b|}{|a-b|}$$

Ej: $a = 0.123467$, $b = 0.123456$

$$K_{rel} = \frac{|0.123467|+|0.123456|}{|0.00011|} = 2244{,}754545$$

¡cuidado!

* Análisis de condición del problema $Ax=b$

Consideremos ahora el problema de resolver un sistema lineal $Ax=b$. Dada una matriz A invertible en $\mathbb{R}^{n\times n}$ y $b\in\mathbb{R}^n$.

Sabemos que el problema tiene solución única ssi $\det(A)\neq 0$. En general, A y b son los datos en la ec. lineal, así, notamos al problema como $(f,(A,b))$, y la función f estaría definida $f: \mathbb{R}^{n\times n}\times\mathbb{R}^n \to \mathbb{R}^n$; tq

$$f(A,b) = A^{-1}b = x.$$

↳ función de los datos

De la def. del cond. de un problema tenemos:

$$\|\hat{x}-x\| = \|\hat{A}^{-1}\hat{b} - A^{-1}b\| \leq K_{rel}\, \|(\hat{A},\hat{b})-(A,b)\|$$

De la def. del cond. de un problema tenemos:

$$\frac{\|\hat{x}-x\|}{\|x\|} = \frac{\|\hat{A}^{-1}\hat{b} - A^{-1}b\|}{\|A^{-1}b\|} \leq K_{rel} \frac{\|(\hat{A},\hat{b})-(A,b)\|}{\|(A,b)\|}$$

Donde la norma del último término está definida en $\mathbb{R}^{n\times n} \times \mathbb{R}^n$. Es así, que es conveniente analizar el problema en 2 etapas:

1. Los errores ocurren en b.
2. Los errores " " A.

1. Caso: errores en b: La función estaría dada por $f: \mathbb{R}^n \to \mathbb{R}^n$

$$b \mapsto A^{-1}b = x.$$

como la función es diferenciable en $\mathbb{R}^n$, tenemos que $K_{abs} = \|f'(b)\| = \|A^{-1}\|$

$$K_{rel} = \frac{K_{abs}\|b\|}{\|A^{-1}b\|} = \frac{\|A^{-1}\|\|b\|}{\|x\|} = \frac{\|A^{-1}\|\|Ax\|}{\|x\|}$$

$$\leq \frac{\|A^{-1}\|\|A\|\|x\|}{\|x\|}$$

$$= \|A^{-1}\|\|A\|$$

con $\|x\| \neq 0$.

2. Caso: errores en A:

$$f: \mathbb{R}^{n\times n} \to \mathbb{R}^n$$

$$A \mapsto A^{-1}b = x$$

$$K_{abs} = \|f'(A)\| = \|Df(A)\|$$

Def: Sean V y W espacios de Banach. Dada la función $f: U \subset V \to W$, decimos que f es diferenciable en el punto $v \in U$, en el sentido fuerte o de Fréchet si existe una aplicación lineal y continua $Df(v): V \to W$, tq.

$$\lim_{\|h\|\to 0} \frac{\|f(v+h) - f(v) - Df(v)h\|}{\|h\|} = 0$$

??

Nota: Si f es Fréchet diferenciable en v

Nota: Si f es Fréchet diferenciable en v en U, su derivada coincide con la derivada direccional $f'(v;h)$ definida por:

$$f'(v;h) = \frac{d}{dt} f(v+th)\Big|_{t=0}$$

$$= \lim_{t\to 0^+} \frac{f(v+th) - f(v)}{t}$$

pero el recíproco no es cierto

Cálculo de la derivada direccional de

$g(A) = A^{-1}$; $g: V \subset \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$
$V = \{B \in \mathbb{R}^{n\times n} \text{ tq } B \text{ es invertible}\}$

$$\frac{d}{dt}(A+tC)^{-1}\Big|_{t=0} = \lim_{t\to 0^+} \frac{(A+tC)^{-1} - A^{-1}}{t}$$

Desarrollo

$$I = (A+tC)(A+tC)^{-1}$$

$$\frac{dI}{dt} = \frac{d}{dt}\left[(A+tC)(A+tC)^{-1}\right]$$

$$0 = \frac{d(A+tC)}{dt}(A+tC)^{-1} + (A+tC)\frac{d(A+tC)^{-1}}{dt}$$

$$= C(A+tC)^{-1} + (A+tC)\frac{d}{dt}(A+tC)^{-1}$$

$$-(A+tC)^{-1} * C(A+tC)^{-1} = \frac{d}{dt}(A+tC)^{-1}$$

Así $$g'(A;C) = \frac{d}{dt}(A+tC)^{-1}\Big|_{t=0} = \lim_{t\to 0^+} \frac{(A+tC)^{-1} - A^{-1}}{t}$$

$$\frac{d}{dt}(A+tC)^{-1}\Big|_{t=} = -(A+tC)^{-1}C(A+tC)^{-1}\Big|_{t=0}$$

$$g'(A;C) = -A^{-1}CA^{-1} = Dg(A)C$$

(g es F-diferenciable)

Regresando a nuestro problema original, debemos encontrar: $Kabs = \|f'(A)\|$

Sin embargo $f = gb \Rightarrow f(A) = A^{-1}b$

Sin embargo $f = gb \Rightarrow f(A) = A^{-1}b$
$= g(A)b$

$$\|Dg(A)\| = \sup_{\|C\|\neq 0} \frac{\|Dg(A)C\|}{\|C\|}$$

$$Df(A) = Dg(A)b$$

$$K_{abs} = \|Df(A)\| = \|Dg(A)b\|$$
$$= \sup_{\|C\|\neq 0} \frac{\|Dg(A)Cb\|}{\|C\|}$$
$$= \sup_{\|C\|\neq 0} \frac{\|-A^{-1}CA^{-1}b\|}{\|C\|}$$
$$= \sup_{\|C\|\neq 0} \frac{\|-A^{-1}Cx\|}{\|C\|}$$
$$\leq \|A^{-1}\|\,\|x\|$$

$$K_{rel} = \frac{K_{abs}\,\|A\|}{\|A^{-1}b\|} \leq \frac{\|A^{-1}\|\,\|x\|\,\|A\|}{\|x\|}$$
$$= \|A^{-1}\|\,\|A\|$$

En las dos etapas se probó

$K_{rel} \leq \|A^{-1}\|\,\|A\|$, independientemente de la norma matricial utilizada.

Def La cantidad

$$\rightarrow K(A) = \|A^{-1}\|\,\|A\|$$

se llama número de condición de la matriz A

Propiedades:

Deber
- $K(A) \geq 1$
- $K(\alpha A) = K(A)$
- $\det(A) = 0 \iff K(A) = \infty$
- $K(A) = \max_{\|x\|=1} \|Ax\| \Big/ \min_{\|x\|=1} \|Ax\|$, generaliza el número de condición para matrices cuadradas.

Resumen: los errores ocurren en b $\rightarrow K_{rel} \leq \|A\|\|A^{-1}\|$
" " " en A

Análisis de condición por componentes

Ej:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix}, \quad 0 < \varepsilon << 1$$

$$K(A)_\infty = \|A^{-1}\|_\infty \|A\|_\infty$$

$$= \varepsilon^{-1}$$

Según el análisis normado, la matriz A es mal condicionada para valores pequeños de ε.

Sin embargo, si resolvemos el sistema:

$$\begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$$\Rightarrow \quad x_1 = b_1$$

$$\varepsilon x_2 = b_2$$

$$\Rightarrow x_2 = \frac{b_2}{\varepsilon}$$

Dado que el sistema de ecuaciones es resuelto sin dificultad, es decir, no se incurre en mayores problemas causados por las pequeñas variaciones del dato ε. Por lo tanto, parece que la matriz es "artificialmente" mal condicionada.

Así, introducimos la sig. def.

Def. La condición absoluta por componentes del problema (f, x) es la constante más pequeña K_{abs}, tq.

$$\|f(\hat{x}) - f(x)\|_\infty \leq K_{abs} \max_i |\hat{x}_i - x_i|$$

para $\hat{x} \to x$

De la misma forma, la condición relativa por componentes es la constante más pequeña K_{rel}, tq.

$$\frac{\|f(\hat{x}) - f(x)\|_\infty}{\|f(x)\|_\infty} \leq K_{rel} \max_i \frac{|\hat{x}_i - x_i|}{|x_i|}$$

para $\hat{x} \to x$

Para un problema f diferenciable, con derivada conocida, caracterizamos el número de condición por componentes como:

$$\rightarrow K_{abs} = \| |f'(x)| \|_\infty$$

$$\rightarrow K_{rel} = \frac{\| |f'(x)| \cdot |x| \|_\infty}{\|f(x)\|_\infty}$$

En este caso, $|f'(x)|$ corresponde a la matriz cuyas componentes son los valores abs. de

En este caso, $|f'(x)|$ corresponde a la matriz cuyas componentes son los valores abs. de las componentes de la matriz $f'(x)$.

Ejercicio: Calcular el número de condición por componentes de problema de multiplicar dos números reales.

$$f: \mathbb{R}^2 \to \mathbb{R}$$
$$\begin{pmatrix} a \\ b \end{pmatrix} \mapsto f(a,b) = a \cdot b$$

$$K_{abs} = \| |f'(x)| \|_\infty = \| |\nabla f^T| \|_\infty$$
$$= \| (|b|, |a|) \|_\infty$$
$$= \sup_{\|x\| \neq 0} \frac{\| (|b|, |a|) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \|_\infty}{\| \begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix} \|_\infty}$$
$$= \sup_{\|x\| = 0} \frac{|x_1|b| + x_2|a||}{\max(|x_1|, |x_2|)} \quad (*)$$

Caso 1 $\max(|x_1|, |x_2|) = |x_1| \Rightarrow |x_2| \leq |x_1|$

De $(*)$
$$\frac{|x_1|b| + x_2|a||}{|x_1|} \leq \frac{|x_1||b| + |x_2||a|}{|x_1|}$$
$$= |b| + \frac{|x_2|}{|x_1|}|a|$$
$$\leq |b| + |a|$$

caso 2 (análogo)
$$\frac{|x_1|b| + x_2|a||}{|x_2|} \leq |b| + |a|$$

$$K_{abs} \leq |b| + |a|$$

$$K_{rel} = \frac{\| (|b|, |a|) \begin{pmatrix} |a| \\ |b| \end{pmatrix} \|_\infty}{|a \cdot b|}$$

$$\leq 2$$

Deber: Análisis por componente para la división.

Análisis de condición por componentes para $Ax = b$.

→ Los errores ocurren en b. $f(b) = A^{-1}b$

$$K_{rel} = \frac{\| |A^{-1}| |b| \|_\infty}{\| A^{-1}b \|_\infty} \qquad |b| = |Ax|$$

$$= \| |A^{-1}| |A| \|_\infty$$

$$\frac{\| A^{-1}b \|_\infty}{\| x \|_\infty} \qquad |b| = |Ax|$$

$$= \frac{\| \, |A^{-1}||b| \, \|_\infty}{\| x \|_\infty}$$

$$\leq \| \, |A| \, \|_\infty \| A^{-1} \|_\infty$$

→ Los errores ocurren en A y b simultáneamente.

En este caso la función $f : \mathbb{R}^{n \times n} \times \mathbb{R}^n \to \mathbb{R}^n$ está dada por $(A, b) \mapsto A^{-1}b$. Haciendo un análisis similar al realizado en el análisis mencionado, tenemos que

$$f'(A,b)(C,d) = -A^{-1}Cx + A^{-1}d$$
$$= -A^{-1}CA^{-1}b + A^{-1}d$$

entonces

$$\frac{\| \, |f'(A,b)(|\epsilon|,|\delta|)| \, \|_\infty}{\|A^{-1}b\|_\infty} \leq \frac{\| \, |A^{-1}||C||x| + |A^{-1}||\delta| \, \|_\infty}{\|x\|_\infty}$$

tomando $(|C|, |d|) = (|A|, |b|)$

$$\kappa_{rel} = \frac{\| \, |f'(A,b)(|A|,|b|)| \, \|_\infty}{\|A^{-1}b\|_\infty}$$

$$\leq 2 \frac{\| \, |A||A^{-1}||x| \, \|_\infty}{\|x\|_\infty}$$

$$\leq 2 \| \, |A||A^{-1}| \, \|_\infty$$

La cantidad $\kappa_C(A) := \| \, |A||A^{-1}| \, \|_\infty$ se denomina condición de Skeel de la matriz A.

Estabilidad de un algoritmo

La estabilidad de un algoritmo se refiere al efecto que produce en las soluciones del problema (f, x) el reemplazo de f por $\hat{f}$, donde $\hat{f}$ contiene todos los errores numéricos cometidos al resolver un problema mediante un método.

→ Nota: Los algoritmos son un conjunto de operaciones aritméticas y lógicas recurrentes que son implementadas mediante operaciones de punto flotante.

Se dice que un algoritmo es **inestable** si se producen grandes errores en los resultados debido a errores que provienen del redondeo

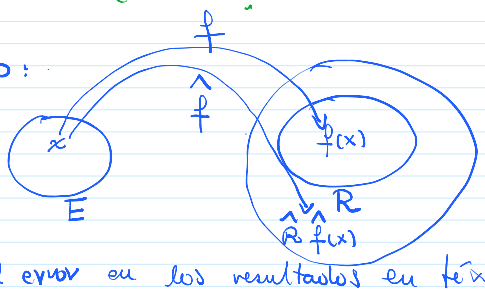
producen grandes errores en los resultados debido a errores que provienen del redondeo en op. de punto flotante, truncamiento, etc.

Estabilidad

↳ Progresivo

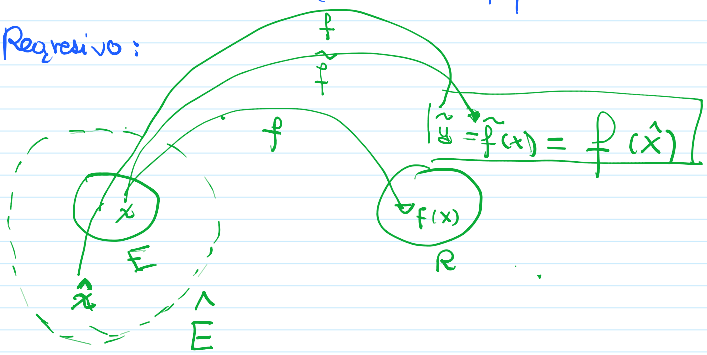
↳ Regresivo

Progresivo:



Estimar el error en los resultados en términos del error en los datos (números de p.f).

Regresivo:



Análisis progresivo:

Para caracterizar un algoritmo $\hat{f}$, se toman en cuenta todas las operaciones elementales $\circ \in \{+,-,\cdot,\div\}$ y su versión de punto flotante tq. satisfacen.

$$\rightarrow \boxed{a \hat{\circ} b = (a \circ b)(1+\varepsilon)}$$

donde $\varepsilon = \varepsilon(a,b)$ tq. $|\varepsilon| \leq \boxed{eps}$

Hay que notar que el remplazo de f por $\hat{f}$ no es único y existe una familia de funciones $\{\hat{f}\}$ que se representa por $\hat{f}$.

Ej: Promedio

$$(a,b) \xrightarrow{f} \frac{a+b}{2}$$

① $(a,b) \xrightarrow{\hat{f}_1} (a \hat{+} b) \hat{\div} 2$

② $(a,b) \xrightarrow{\hat{f}_2} (a \hat{\div} 2) + (b \hat{\div} 2)$

Obs: El análisis progresivo consiste en determinar cuán diferentes son $R = f(E)$ y $\hat{R} = \hat{f}(E)$

Obs: El análisis progresivo consiste en aproximar ecuán diferentes son $R = f(E)$ y $\hat{R} = \hat{f}(E)$

Def Sean E el conjunto de los datos y $\hat{f}$ la evaluación en punto flotante de un algoritmo para resolver un problema (f, x) que tiene condición relativo normado: $K_{rel} := K_f$. El indicador de estabilidad del análisis progresivo normado es el valor $\sigma_f \geq 0$ más pequeño tq.

$$\frac{\|\hat{f}(x) - f(x)\|}{\|f(x)\|} \leq \sigma_f K_f \, eps$$

σ_f: indicador de estabilidad

con $x \in E$, $eps \to 0$

Se dice que un algoritmo $\hat{f}$ es estable en el sentido del análisis progresivo si σ_f es menor o igual que el número de operaciones que se efectúan al evaluar $\hat{f}$.

Ej: La evaluación en punto flotante $\{\hat{+}, \hat{-}, \hat{\cdot}, \hat{\div}\}$ de las operaciones elementales verifican:

$$\frac{|a \circ b - a \hat{\circ} b|}{|a \circ b|} = \frac{|a \circ b - a \circ b(1+\varepsilon)|}{|a \circ b|}$$

$$= \frac{|a \circ b||\varepsilon|}{|a \circ b|} = |\varepsilon| \leq eps$$

$$\Rightarrow \frac{|a \circ b - a \hat{\circ} b|}{|a \circ b|} \leq eps$$

Así, $\boxed{\sigma_K \leq 1}$

Teorema Un problema (f, x) con indicador de estabilidad σ_f, el cual se subdivide en subproblemas $(h, g(x))$ y (g, x) con $f = h \circ g$ $(f(x) = h(g(x)))$ con indicadores de estabilidad σ_h y σ_g y números de condición relativos normados K_h y K_g resp., entonces

$$\sigma_f K_f \leq \sigma_h K_h + \sigma_g K_g K_h$$

Demos-

$$\|f(x)-\hat{f}(x)\| = \|h(g(x)) - \hat{h}(\hat{g}(x))\|$$

$$= \|h(g(x)) - h(\hat{g}(x)) + h(\hat{g}(x)) - \hat{h}(\hat{g}(x))\|$$

$$\leq \|h(g(x)) - h(\hat{g}(x))\| + \|h(\hat{g}(x)) - \hat{h}(\hat{g}(x))\|$$

$$\leq K_h \|h(g(x))\| \frac{\|g(x)-\hat{g}(x)\|}{\|g(x)\|} + G_h K_h \text{eps} \|h(\hat{g}(x))\|$$

$$\leq K_h \|h(g(x))\| G_g K_g \text{eps} + G_u K_u \text{eps} \|h(\hat{g}(x))\|$$

$$\leq \|h(g(x))\| (K_h G_g K_g \text{eps} + G_u K_h \text{eps})$$

$$= \|f(x)\| (G_g K_g K_h + G_u K_u) \text{eps}$$

$$\frac{\|f(x)-\hat{f}(x)\|}{\|f(x)\|} \leq (G_g K_g K_h + G_u K_u) \text{eps}$$

$$\Rightarrow G_f K_f \leq G_g K_g K_h + G_u K_u$$

para eps $\to 0$. ∎

Ej: Análisis de estabilidad del algoritmo de la suma de n números reales.

$$S_n : \mathbb{R}^n \to \mathbb{R} \quad , \quad S_n(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$$

Definimos S_n como la composición de dos funciones recursivas:

$$\alpha_n : \mathbb{R}^n \to \mathbb{R}^{n-1} : \alpha_n(x_1, \dots, x_n) = (x_1 + x_2, x_3, x_4, \dots, x_n)$$

la cual nos permite expresar S_n recursivamente como la composición de:

$$S_n = S_{n-1} \circ \alpha_n \quad , \quad n \geq 2$$

Obs: ① $G_{\alpha_n} = G_+$

② $K_{\alpha_n} = K_+$

③ $S_n = S_{n-1} \circ \alpha_n \Rightarrow G_n K_n \leq K_{n-1} (G_{n-1} + G_+ K_+)$

↳ $K_+ = \dfrac{|a|+|b|}{|a+b|}$

$$1 = \frac{|a+b|}{|a+b|} \leq \frac{|a|+|b|}{|a+b|} = K_+$$

↳ $K_{n-1} = \dfrac{|x_1+x_2| + \sum_{i=3}^{n} |x_i|}{\left|\sum_{i=1}^{n} x_i\right|}$

$$\leq \frac{\sum_{i=1}^{n} |x_i|}{\left|\sum_{i=1}^{n} x_i\right|} = K_n$$

↳ $G_+ K_+ \leq 1$

$$\sigma_n K_n \le K_{n-1}(\sigma_{n-1} + \sigma_+ K_+)$$
$$\le K_n(\sigma_{n-1} + \sigma_+ K_+)$$

$$\Rightarrow \sigma_n \le \sigma_{n-1} + \sigma_+ K_+$$
$$\le \sigma_{n-1} + 1$$

Tomemos σ_2, $\Rightarrow \sigma_2 = \sigma_+ \le \frac{1}{K_+} \le 1$

$n=3 \Rightarrow \sigma_3 \le \sigma_2 + 1$
≤ 2

$n=4 \Rightarrow \sigma_4 \le \sigma_3 + 1$
≤ 3

$\vdots$

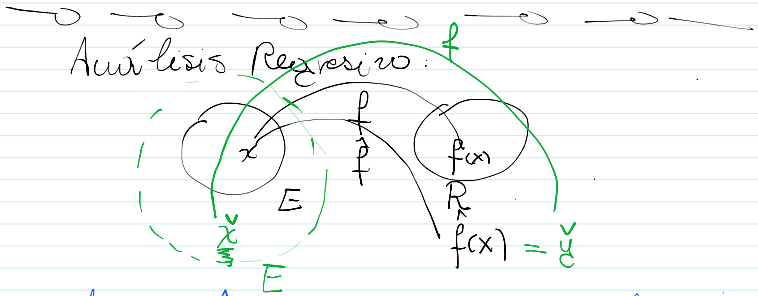
$$\sigma_n \le n-1$$ ✓

↳ Es estable :)

Def Sea $\hat{f}$ la evaluación en punto flotante de un algoritmo para resolver un problema (f, X) que tiene condición por componente $x_+ := \kappa_{rel}$. El indicador de estabilidad del A.P. comp. a comp. es el valor $\sigma_f \ge 0$ más pequeño tq

$$\max_i \frac{|\hat{f}_i(x) - f_i(x)|}{|f_i(x)|} \le \sigma_f \kappa_f \, eps$$

para $eps \to 0$.



Def Sea $\hat{f}$ la evaluación en punto flotante de un algoritmo para resolver el problema (f,x) donde, $f: E \to R$. El error de estabilidad del análisis regresivo normado es el menor número $\eta_f \ge 0$ tq para cada $x \in E$ existe $\check{x}$ tq

$$f(\check{x}) = \hat{f}(x)$$

y se cumple

$$\frac{\|x - \check{x}\|}{\|x\|} \le \eta_f \quad \text{para } eps \to 0$$

y se cumple:

$$\frac{\|x - \check{x}\|}{\|x\|} \dot{\leq} \eta_f \quad \text{para } eps \to 0$$

De igual manera, se define el error de estabilidad del AR por comp. como el menor número $\eta_f \geq 0$ tq, para cada $x \in E$, existe $\check{x}$ tq

$$\max_i \frac{|x_i - \check{x}_i|}{|x_i|} \dot{\leq} \eta_f$$

El análisis regresivo tiene la ventaja de que no depende de la condición del problema. El algoritmo $\hat{f}$ se dice estable en el sentido del AR con respecto a un error relativo $\delta > 0$, ssi $\eta_f \leq \delta$. El indicador del AR se define por

$$G_r = \frac{\eta_f}{eps}, \quad \text{para } eps \to 0$$

Los indicadores de estabilidad del AP y AR verifican para un problema (f, x):

$$\boxed{G_f \leq G_r}$$

pues se tiene que para cada $x \in E$, existe $\check{x}$ tq $\hat{f}(x) = f(\check{x})$ y se verifica

$$\frac{\|x - \check{x}\|}{\|x\|} \dot{\leq} \eta_f$$

$$\frac{\|\hat{f}(x) - f(x)\|}{\|f(x)\|} = \frac{\|f(\check{x}) - f(x)\|}{\|f(x)\|}$$

$$\dot{\leq} K_{rel} \frac{\|x - \check{x}\|}{\|x\|}$$

$$\dot{\leq} K_{rel}\, \eta_f$$

$$= K_{rel}\, G_r\, eps \quad (*) \quad \text{para } eps \to 0$$

Sabemos G_f es el número más pequeño que satisface $(*)$. Por lo tanto:

$$\boxed{G_f} \leq \boxed{G_r}$$

Así, si un problema es estable en el sentido de AR, también lo es en el AP.

Ej: Análisis regresivo para el producto escalar en $\mathbb{R}^n$.

$$\langle \cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$$
$$(x,y) \longmapsto \langle x,y\rangle = \sum_{i=1}^{n} x_i y_i$$

Esta función se puede escribir de la siguiente manera:

$$\langle x,y\rangle = x_n y_n + \langle \bar{x},\bar{y}\rangle$$

donde $\bar{x} = (x_1,\dots,x_{n-1})$

$\bar{y} = (y_1,\dots,y_{n-1})$.

Así, vamos a probar que existe $\check{x} \in \mathbb{R}^n$ tq. $\langle x,y\rangle_{fl} = \langle \check{x},y\rangle$

y $$\frac{|x_i - \check{x}_i|}{|x_i|} \leq n\,eps, \quad i = 1,\dots,n.$$
$\eta_f \leq n\,eps$

Procedemos por inducción:

1⋆ $n = 1$ $\quad \hat{x}\cdot y = (x\cdot y)(1+\varepsilon)$, $|\varepsilon| \leq eps$

$= x(1+\varepsilon)\cdot y$

$= \check{x}\cdot y$.

Así, definimos $\check{x} = (1+\varepsilon)x$ que satisface

$$\hat{x\cdot y} = \check{x}\cdot y \quad y$$
$$\frac{|x - (1+\varepsilon)x|}{|x|} = \frac{|\varepsilon|\,|x|}{|x|} \leq 1\,eps.$$

2⋆: Suponemos que la propiedad se satisface para $(n-1)$ y probamos para n. Entonces, nuestra hipótesis inductiva se escribe de la siguiente manera:

HI) Existe $\check{z} \in \mathbb{R}^{n-1}$ tq.

$$\langle \bar{x},\bar{y}\rangle_{fl} = \langle \check{z},\bar{y}\rangle \quad y$$
$$\frac{|\check{z}_i - \bar{x}_i|}{|\bar{x}_i|} \leq (n-1)\,eps.$$

En general para $n>1$,

$\langle x,y\rangle_{fl} = [x_n y_n(1+\varepsilon_1) \oplus \langle \bar{x},\bar{y}\rangle_{fl}]\,(1+\varepsilon_2)$, $|\varepsilon_1|,|\varepsilon_2| \leq eps$

$= x_n(1+\varepsilon_1)(1+\varepsilon_2)\,y_n + \langle \check{z},y\rangle\,(1+\varepsilon_2)$

$= \check{x}_n y_n + \langle \check{z}(1+\varepsilon_2), y\rangle$

$= \check{x}_n y_n + \langle (\check{x}_1, \dots, \check{x}_{n-1}), y \rangle = \langle \check{x}, y \rangle$

con $\check{x}_i = \check{z}_i (1+\varepsilon_2)$, $i = 1, \dots, n-1$

Aquí $\check{x} = (\check{x}_1, \dots, \check{x}_n)$.

Nos queda por probar que

$$\frac{|\check{x}_i - x_i|}{|x_i|} \leq n \, eps, \quad \text{para todo } i = 1, \dots, n$$

por lo tanto:

$$\begin{aligned} |x_i - \check{x}_i| &= |x_i - \check{z}_i (1+\varepsilon_2)| \quad \text{para } i = 1, \dots, n-1 \\ &\leq |x_i - \check{z}_i| + |\varepsilon_2| \, |\check{z}_i| \\ &\leq |x_i - \check{z}_i| + eps \, |\check{z}_i| \\ &\leq (n-1) \, eps \, |x_i| + eps \, |\check{z}_i| \\ &\leq (n-1) \, eps \, |x_i| + eps |x_i| + O(eps) \quad \text{para } eps \to 0 \\ &\doteq n \, eps \, |x_i| \end{aligned}$$

$$\Rightarrow \quad \frac{|x_i - \check{x}_i|}{|x_i|} \leq n \, eps, \quad i = 1, \dots, n-1$$

Para $i = n$

$$|x_n - x_n (1+\varepsilon_1)(1+\varepsilon_2)| \leq 2 \, eps$$

Por lo tanto,

$$\max_{1 \leq i \leq n} \frac{|x_i - \check{x}_i|}{|x_i|} \leq n \, eps,$$

Así $\quad G_r = \frac{\eta_f}{eps} \Rightarrow \eta_f \leq n \, eps$

Resumen:

- Condicionamiento (Problema(f, x)) caso dif $\kappa_{abs} = \|f'(x)\|$, κ_{rel}
- Estabilidad ($\tilde{G}$)
  - ΔP → Diferencia entre P y $\hat{P}$ (Si se necesita κ_{rel})
  - ΔR → Análisis respecto a los datos.

Clase de ejercicios

- Cond de un problema → κ_{abs} / κ_{rel}
- Estabilidad $\tilde{G}_{r,f}$ → ΔR / ΔP

(1) ¿Cuál es el número de condición relativo normado de $f(x)=e^x$? ¿Cuándo se vuelve $f(x)$ mal condicionado?

$$f'(x) = e^x$$

$$K_{abs} = \|f'(x)\| = \sup_{y \neq 0} \frac{\|f'(x)\, y\|}{\|y\|}$$

$$= \sup_{y \neq 0} \frac{|e^x|\,\|y\|}{|y|}$$

$$= |e^x|$$

$$K_{rel} = \frac{|e^x|\,|x|}{|e^x|} = |x|$$

$\Rightarrow$ El problema está mal condicionado

$x \to \infty$

$x \to -\infty$

(2) Dado un vector u, la proyección ortogonal de un vector v sobre v está dada por

$$f(v) = \frac{\langle u, v\rangle}{\langle u, u\rangle} \cdot u$$

Muestra que $K_{rel} \leq \sec\theta$.

Nota: $\cos\theta = \dfrac{\langle u, v\rangle}{\|u\|\,\|v\|}$

$$f : \mathbb{R}^u \to \mathbb{R}^u$$

$$v \longmapsto \frac{\langle u, v\rangle}{\|u\|^2} \cdot u = \frac{u^T v}{\|u\|^2} \cdot u$$

$$f'(v) = \frac{u u^T}{\|u\|^2}$$

$$K_{abs} = \|f'(v)\| = \sup_{\|x\| \leq 1} \left\| \frac{u u^T}{\|u\|^2} \cdot x \right\|$$

$$= \sup_{\|x\| \leq 1} \left\| \frac{u \langle x, u\rangle}{\|u\|^2} \right\|$$

Pues:

$$= 1$$

$$\left\| \frac{u}{\|u\|^2} \langle x, u\rangle \right\| = \left\| \frac{u}{\|u\|^2} \right\| |\langle x, u\rangle|$$

$$\leq \left\| \frac{u}{\|u\|^2} \right\| \|x\|\,\|u\|$$

$$\leq \left\| \frac{u}{\|u\|^2} \right\| \|x\| \|u\|$$

$$\leq \frac{\|u\|}{\|u\|^2} \cdot \|u\| = 1$$

$$K_{abs} = 1$$

Ahora, calculamos K_{rel}:

$$K_{rel} = \frac{1 \cdot \|v\|}{\left\| \frac{\langle u, v \rangle u}{\|u\|^2} \right\|} = \frac{\|v\|}{\frac{|\langle u, v \rangle| \|u\|}{\|u\|^2}}$$

$$= \frac{\|v\| \|u\|}{|\langle u, v \rangle|}$$

$\langle u, v \rangle \leq |\langle u, v \rangle|$

$\frac{1}{|\langle u, v \rangle|} \leq \frac{1}{\langle u, v \rangle}$

$$\leq \sec(\theta)$$

(3) Sean las funciones diferenciables f, g y h definidas en $\mathbb{R}$ con valores en $\mathbb{R}$, tq $f = h \circ g$. Demuestre que

$$C_f \leq \frac{C_h}{K_g} + C_g.$$

$$K_f = \frac{|f'(x)| |x|}{|f(x)|} = \frac{|h'(g(x))\, g'(x)|\, |x|}{|h(g(x))|}$$

$$= \frac{|h'(g(x))|\, |g'(x)|\, |x|\, |g(x)|}{|h(g(x))|\, |g(x)|}$$

$$= K_h \cdot K_g$$

Aplicando este resultado de Teo visto en clase tenemos:

$$C_f \leq \frac{C_g K_g K_h}{K_f} + \frac{C_h K_h}{K_f}$$

$$= C_g + \frac{C_h}{K_g}.$$

(4) Determine si el problema

④ Determine si el problema

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}$$
$$x \longmapsto x^T x = \sum_{i=1}^{n} x_i^2 = (x_1^2 + x_2^2 + \dots + x_n^2)$$

es bien condicionado en el sentido del análisis numérico en norma euclideana.

$$f'(x) = 2\,(x_1 \;\; x_2 \;\; x_n) = 2x^T$$

$$K_{abs} = \|f'(x)\| = \sup_{y \neq 0} \frac{\|2x^T y\|_2}{\|y\|_2} = \frac{2\,|\langle x, y\rangle|}{\|y\|} \leq \frac{2\,\|x\|\|y\|_2}{\|y\|_2} = 2\|x\|_2$$

$$K_{rel} = \frac{2\|x\|\,\|x\|}{x^T x} = \frac{2\|x\|\|x\|}{\|x\|^2} = 2$$

⑤ Análisis por componentes del producto escalar:

$$f : \mathbb{R}^{2n} \longrightarrow \mathbb{R}^1$$
$$(x, y) \longmapsto \langle x, y\rangle = x^T y = \sum_{i=1}^{n} x_i y_i$$

$$K_{abs} = \| \, |f'(x)| \, \|_\infty$$

$$f'(x, y) = (y^T, x^T) \in \mathbb{R}^{2n} =$$

$$K_{abs} = \| \, |(y^T, x^T)| \, \|_\infty$$
$$= \| \, (|y^T|, |x^T|) \, \|_\infty$$
$$= \| \, (|y_1|, |y_2|, \dots, |y_n|, |x_1|, \dots, |x_n|) \, \|_\infty$$

$$K_{rel} = \frac{\| \, (|y^T|, |x^T|) \cdot (|x|, |y|) \, \|_\infty}{\| \quad \|\langle x, y\rangle\|_\infty}$$

$$\begin{pmatrix} |x_1| \\ |x_2| \end{pmatrix}$$

$$= \frac{\| \langle x, y \rangle \|_\infty}{\| (|y_1|, |y_2|, \ldots, |y_n|, |x_1|, \ldots, |x_n|) \cdot \begin{pmatrix} |x_1| \\ |x_2| \\ \vdots \\ |x_n| \\ |y_1| \\ \vdots \\ |y_n| \end{pmatrix} \|_\infty} \Big/ \| \langle$$

$$= \frac{2 \langle |x|, |y| \rangle}{|\langle x, y \rangle|}$$

Métodos directos para la resolución de sistemas lineales

$$Ax = b$$

Este problema tiene solución única sii $\det(A) \neq 0$, con $A \in \mathbb{R}^{n \times n}$ y $b \in \mathbb{R}^n$.

La solución del sistema $Ax = b$ está dada por $x = A^{-1} b$. Sin embargo, calcular la inversa de A requiere demasiadas operaciones.

Ej: $\det(A)$ usando la fórmula de Laplace tenemos:

$$\det(A) = \sum_{i=1}^{n} (-1)^{i+1} a_{1i} \det(A_{1i})$$

2^n operaciones o flops

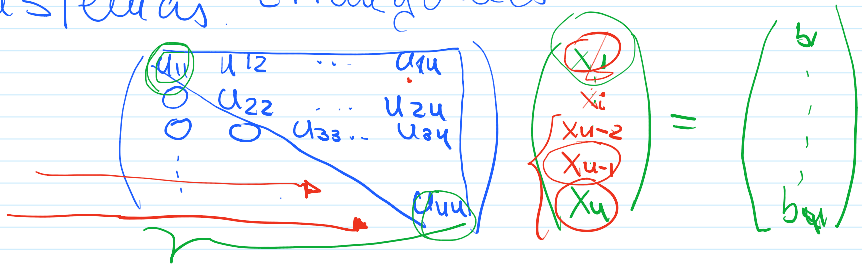
Idea: No calcular A^{-1}

$\rightarrow$ ~~$x = A^{-1} b$~~ Muy caro

Sistemas triangulares:

$\|x, y\|_\infty$

Sistemas triangulares:



$$Ux = b$$

$$u_{nn} x_n = b_n \Rightarrow x_n = \frac{b_n}{u_{nn}}$$

$$u_{n-1\,n-1} x_{n-1} + u_{n-1\,n} x_n = b_{n-1}$$

$$\Rightarrow x_{n-1} = \frac{b_{n-1} - u_{n-1\,n} \cdot x_n}{u_{n-1\,n-1}}$$

$$\vdots$$

$$x_i = \frac{1}{u_{ii}} \left(b_i - \sum_{k=i+1}^{n} u_{ik} x_k \right)$$

Algoritmo para la resolución de sistemas triangulares:

Sus Rég. Input: $U \in \mathbb{R}^{n \times n}$ ($\triangle$), $b \in \mathbb{R}^n$

Output: $x \in \mathbb{R}^n$, $Ux = b$

for $i = n:1$

$$x_i = \frac{1}{u_{ii}} \left(b_i - \sum_{k=i+1}^{n} x_k u_{ik} \right)$$

end

Sus Prog. Input: $L \in \mathbb{R}^{n \times n}$ ($\triangle$), $b \in \mathbb{R}^n$

Output: $x \in \mathbb{R}^n$, $Lx = b$

for $i = 1:n$

$$x_i = \frac{1}{l_{ii}} \left(b_i - \sum_{k=1}^{i-1} x_k l_{ik} \right)$$

end

Obs: La complejidad de un algoritmo se define como el orden de operaciones

UDS: La complejidad de un algoritmo se define como el orden de operaciones que se efectúan para resolver un problema determinado.

Orden: se refiere al término que gobierna el número de operaciones.

Análisis del Alg. de sust. Regresiva:

$$x_i = \frac{1}{u_{ii}} \left(b_i \ominus \sum_{k=i+1}^{n} x_k u_{ik} \right)$$

Componente	Mult.	Divis.	Sumas/Restas
x_n	0	1	0
x_{n-1}	1	1	1
x_{n-2}	2	1	2
x_{n-3}	3	1	3
⋮	⋮	⋮	⋮
x_1	$n-1$	1	$n-1$

Luego tenemos:

* n divisiones
* $\sum_{i=1}^{n-1} i$ multiplicaciones
* $\sum_{i=1}^{n-1} i$ sumas y restas

$$\text{Total} = n + 2\sum_{i=1}^{n-1} i = n^2 = O(n^2)$$

$$\sum_{i=1}^{n} i \quad \frac{n(n+1)}{2}$$

$f(n) = O(n^2)$ ssi existe una constante real positiva M y un real n_0 tq

$$f(n) \leq M n^2 \quad \forall n \geq n_0$$

$Ax = b$

$A = L \cdot U = (\Delta)(\nabla)$

$LUx = b$ ($Ux = y$)

$Ly = b$ ① $\Rightarrow y$

$Ux = y$ ② $\Rightarrow x$

Definición: Se dice que $A \in \mathbb{R}^n \times \mathbb{R}^n$ tiene factorización LU si existen matrices L (triang inf Δ) y U triangular superior tq $A = LU$

Ej: $A = \begin{pmatrix} 1 & 2 & 2 \\ 1/2 & 2 & 0 \\ 0 & -2 & -3 \end{pmatrix}$, $b = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

Usaremos el algoritmo de Eliminación Gaussiana.

$A^{(1)} = \left(\begin{array}{ccc|c} 1 & 2 & 2 & 1 \\ 1/2 & 2 & 0 & 0 \\ 0 & -2 & -3 & 0 \end{array}\right)$ $\quad F_2^{(2)} = F_2^{(1)} - \frac{1}{2} F_1^{(1)} \Rightarrow L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

$A^{(2)} = \left(\begin{array}{ccc|c} 1 & 2 & 2 & 1 \\ 0 & 1 & -1 & -1/2 \\ 0 & -2 & -3 & 0 \end{array}\right)$ $\quad F_3^{(3)} = F_3^{(2)} + 2 F_2^{(2)} \Rightarrow L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$

$A^{(3)} = \left(\begin{array}{ccc|c} 1 & 2 & 2 & 1 \\ 0 & 1 & -1 & -1/2 \\ 0 & 0 & -5 & -1 \end{array}\right)$

$A^{(2)} = L_1 A^{(1)}$

$A^{(3)} = L_2 A^{(2)}$

$A^{(3)} = L_2 A^{(2)} = L_2 L_1 A^{(1)} = \underbrace{L_2 L_1} A$

$L_2^{-1} A^{(3)} = L_1 A$

$L_1^{-1} L_2^{-1} A^{(3)} = A$

$\underbrace{L_1^{-1} L_2^{-1}}_{L} U = A$

$\Rightarrow A = L \cdot U$

Fecha: 21 de diciembre

$$\rightarrow \quad A = \underset{\triangle}{L} \cdot \underset{\nabla}{U}$$

↳ Eliminación Gaussiana.

En nuestro ejemplo de la clase anterior teníamos que

$$A^{(3)} = L_2 A^{(2)} = L_2 L_1 A^{(1)} = L_2 L_1 A$$

$$\Rightarrow A = \underbrace{L_1^{-1} L_2^{-1}}_{=L} \underbrace{A^{(3)}}_{U}$$

Nota: Las matrices L_i se denominan matrices de Frobenius y son fácilmente invertibles. Su inversa se obtiene al cambiar de signo a los elementos bajo la diagonal.

$$L_1^{-1} = \begin{pmatrix} 1 & & \\ -1/2 & 1 & \\ & & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & & \\ \frac{1}{2} & 1 & \\ & & 1 \end{pmatrix}$$

$$L_2^{-1} = \begin{pmatrix} 1 & & \\ & 1 & \\ & 2 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & & \\ & 1 & \\ & -2 & 1 \end{pmatrix}$$

Así $L = L_2^{-1} L_1^{-1} = \begin{pmatrix} 1 & & \\ \frac{1}{2} & 1 & \\ & -2 & 1 \end{pmatrix}$

Ahora vamos a deducir el proceso de EG. para una matriz general:

$\triangle$ $a_{11}^{(1)} \neq 0$!

$$A^{(1)} = A = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & \\ \vdots & & \ddots & \\ a_{n1}^{(1)} & \cdots & & a_{nn}^{(1)} \end{pmatrix}$$

$$\boxed{l_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}}$$

$$F_i^{(2)} = F_i^{(1)} - l_{i1} F_1^{(1)}$$

$i = 2, \dots, n$

$$\Rightarrow L_1 \begin{pmatrix} 1 & & \\ -l_{21} & \ddots & \\ -l_{31} & & \\ \vdots & & \\ -l_{n1} & & 1 \end{pmatrix}$$

$\triangledown$ $a_{\ldots}^{(2)}$, ∇

$\Rightarrow L_1 = \begin{pmatrix} 1 & & \\ -l_{21} & \ddots & \\ -l_{n1} & & 1 \end{pmatrix}$

$a_{22}^{(2)} \neq 0$

$A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}$ $(n-1)\times(n-1)$

$l_{i2} = \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}}$

$F_i^{(3)} = F_i^{(2)} - l_{i2} F_2^{(2)}$

$i = 3, \ldots, n$

$\Rightarrow L_2 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & -l_{32} & \ddots & \\ & -l_{n2} & & 1 \end{pmatrix}$

$\vdots$

k-iteración:

$a_{kk}^{(k)} = 0$

$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & & a_{1n}^{(1)} \\ & a_{22}^{(2)} & & a_{2n}^{(2)} \\ & & a_{kk}^{(k)} & \cdots \; a_{kn}^{(k)} \\ & & a_{k+1,k} & a_{k+1,n} \\ & & a_{nk}^{(k)} & a_{nn}^{(k)} \end{pmatrix}$

$l_{ik} = \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$

$F_i^{(k+1)} = F_i^{(k)} - l_{ik} F_k^{(k)}$

$i = k+1, \ldots, n$

$L_k = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & 1 & & \\ & -l_{k+1,k} & \ddots & \\ & -l_{nk} & & 1 \end{pmatrix}$

En este proceso generamos una sucesión de matrices:

$$A = A^{(1)} \to A^{(2)} \to A^{(3)} \cdots \to A^{(n)} = U$$

Así $A^{k+1} = L_k A^k$, $k = 1, \ldots, n-1$

De la misma manera que el ejemplo anterior las inversas de L_k, $k = 1, \ldots, n-1$ se obtienen cambiando de signo a los términos

cambiando de signo a los términos l_{ik}, $i = k+1, \dots, n$, $k = 1, \dots, n-1$.

$$U = A^{(n)} = L_{n-1} L_{n-2} \cdots L_2 L_1 A,$$

despejando A tenemos:

$$L = L_1^{-1} L_2^{-1} \cdots L_n^{-1}$$

con L una matriz triangular inferior (Δ) con 1's en la diagonal.

Obs $$\det(A) = \det(L)\det(U)$$
$$= \det(U) = \prod_{k=1}^{n} a_{kk}^{(k)}$$

Teo: Sea $A \in \mathbb{R}^{n \times n}$ una matriz no singular tq los pivotes en el algoritmo de E.G no se anulan, entonces existe $L \in \mathbb{R}^{n \times n}$ triang. inf (Δ) con los elementos de su diagonal iguales a 1 y $U \in \mathbb{R}^{n \times n}$ triang. sup (∇) tq.

$$A = LU$$

Además, L y U son únicas.

Obs:
- No toda matriz invertible tiene Fact. LU
- Existen matrices sing. que tiene F. LU
- La fact. LU no es única.

Clases de matrices para las cuales se puede aplicar el método de E.G:

- Matrices diagonal dominantes por filas: $|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \forall i = 1, \dots, n$

columnas

- Matrices simétricas definidas positivas

Teo: Sea $A \in \mathbb{R}^{n \times n}$. La fact. LU de A con $l_{ii} = 1$ para $i = 1, \dots, n$ existe y es única ssi las submatrices principales A_i de A de orden $i = 1, \dots, n-1$ son no singulares.

Peder: Ver la demost. Quarteroni Th 3.4

Algoritmo de E.G para encontrar Fact. LU

Input: $A \in \mathbb{R}^{n \times n}$

Output: $L \in \mathbb{R}^{n \times n}$ (Δ) y $U \in \mathbb{R}^{n \times n}$ (∇)

```
for k : 1 : n-1
    for i = k+1 : n
        l_ik = a_ik / a_kk   ←
        for j = k : n
            a_ij = a_ij - l_ik a_kj   ←
        end
    end
end
```

Análisis de Complejidad:

Cada l_{ik} se calcula con una división, por lo tanto, para calcular TODOS los l_{ik} se tiene

$$\sum_{k=1}^{n-1} k = \frac{n(n-1)}{2} = \frac{n^2 - n}{2} \quad (1)$$

Ahora, para las operaciones de fila tenemos que en cada paso k, modificamos $(n-k) \times (n-k)$ entradas de la matriz, es decir, desde $(n-1) \times (n-1)$ en la 1ra etapa hasta (1×1) en la última.

Cada operación se realiza con una suma y una multiplicación. Entonces, el costo es:

$$2\sum_{k=1}^{(n-1)} k^2 = \frac{2}{3}n^3 + \frac{n^2}{3} - \frac{n}{3} \qquad (2)$$

De ① y ②

$$\frac{2}{3}n^3 + \frac{n^2}{3} - \frac{n}{3} + \frac{n^2}{2} - \frac{n}{2} = O\left(\frac{2}{3}n^3\right)$$

Para resolver $Ax = b$

↳ fact. LU $\quad O\left(\frac{2}{3}n^3\right)$

↳ $L\underbrace{Ux}_{y} = b$ $\quad Ly = b \quad O(n^2)$

$\qquad\qquad\qquad\qquad Ux = y \quad O(n^2)$

Estrategias de pivoteo

$$A = \left(\begin{array}{cc|c} \boxed{1\times10^{-4}} & 1 & 1 \\ 1 & 1 & 2 \end{array}\right), \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

$$F_2^{(2)} = F_2^{(1)} - \boxed{1\times10^{4}}\, F_1^{(1)}$$

$$A^{(2)} = \left(\begin{array}{cc|c} 1\times10^{-4} & 1 & 1 \\ 0 & 1-1\times10^{4} & 2-10^{4} \end{array}\right)$$

$$= \left(\begin{array}{cc|c} \boxed{1\times10^{-4}} & 1 & 1 \\ 0 & -9999 & -9998 \end{array}\right)$$

$$1\times10^{-4}x_1 + x_2 = 1$$

$$-9999x_2 = -9998$$

$$x_2 = 0{,}999 \approx 1$$

$$\Rightarrow 1\times10^{-4}x_1 + 1 = 1$$

$$\Rightarrow x_1 = 0$$

Ahora, intercambiamos las filas:

$$\tilde{A} = \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 1\times10^{-4} & 1 & 1 \end{array}\right),$$

$$F_2^{(2)} = F_2^{(1)} - (1\times10^{-4})F_1^{(1)}$$

$$\tilde{A}^{(2)} = \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1-1\times10^{-4} & 1-2\times10^{-4} \end{array}\right)$$

Truncado a 3 decimales

$$\tilde{A}^{(2)} = \left(\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 1 \end{array}\right) \Rightarrow x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

- Los efectos de redondeo suelen diferencias en los sol cuando calculamos los coef de la matriz de Frobenius.
- La idea es mantener los $|l_{ij}|$ lo más pequeños posible.
- Procedimiento: Intercambiar las filas

* Pivoteo parcial por filas:

En el k-ésimo paso, buscamos el índice i^* tq

$$|a_{i^*k}^{(k)}| = \max_{k\le i\le n} |a_{ik}^{(k)}|$$

$$\Rightarrow |l_{ik}| = \left|\frac{a_{ik}^{(k)}}{a_{i^*k}^{(k)}}\right| \le 1$$

* Problema: $\boxed{A \neq LU}$

Al realizar el intercambio de filas, se obtiene una matriz:

$$\tilde{A}^{(k)} = P_k A^{(k)}$$

donde P_k es la matriz de permutación que intercambia la fila k con la i^* en el paso k y se obtiene intercambiando en la matriz identidad la fila k por i^*.

Nota: Al realizar el intercambio de filas para obtener el mejor pivote destruimos la estructura de la matriz A y $\boxed{A \neq LU}$ en este caso:

$$A^{k-1} = L_k \tilde{A}^{(k)} = L_k P_k A^{(k)}, \quad k = 1, \dots, n-1$$

$$U = A^{(n)} = \underbrace{L_{n-1} P_{n-1} \dots L_2 P_2 L_1 P_1}_{\tilde{L}^{-1}} A^{(1)}$$

$$\Rightarrow \tilde{L} U = A$$

$\tilde{L}$ no es triangular inf. (Δ).

Podemos expresar $\boxed{A = PLU}$, donde L sí es Δ y $P = P_1 P_2 \dots P_{n-1}$.

Ej:

$$U = A^{(5)} = L_4 P_4 L_3 P_3 L_2 P_2 L_1 P_1 A^{(1)}$$

$$A = A^{(1)} = P_1 L_1^{-1} P_2 L_2^{-1} P_3 L_3^{-1} P_4 L_4^{-1} U$$

$$= \underbrace{P_1 (P_2 P_3 P_4)}_{P} \underbrace{(P_4 P_3 P_2 L_1^{-1} P_2 P_3 P_4)(P_4 P_3 L_2^{-1} P_3 P_4)(P_4 L_3^{-1} P_4) L_4^{-1}}_{L \text{ (sí es } \Delta)} \underbrace{U}_{U}$$

$$\Rightarrow A = PLU$$

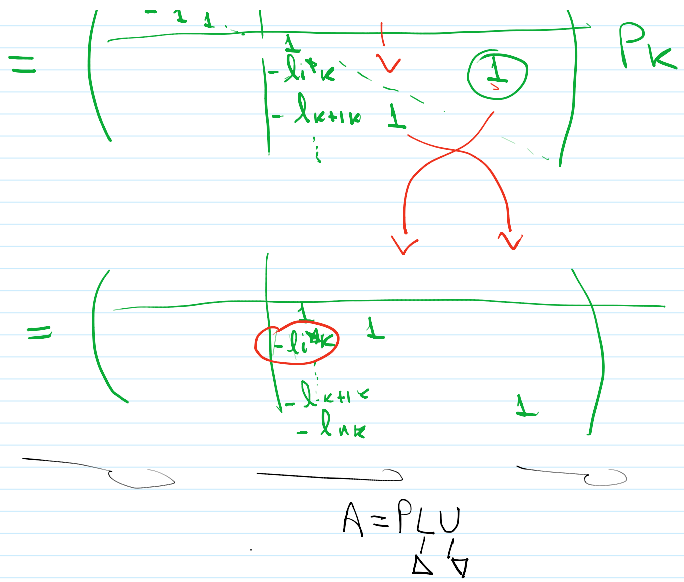
$A = PLU$ ⇒ $AX = PLUX$

$PA = LU$

Notemos que $P_r L_k^{-1} P_r$, $r > k$ nos da:

$$P_k \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & -l_{k+1\,k} & 1 & \\ & & & -l_{i\,k} & & 1 \\ & & & -l_{n\,k} & & & 1 \end{pmatrix} P_k$$

$$= \begin{pmatrix} 1 & 1 & 1 & \dots & & & \\ & & & & 1 & & \\ & & & & -l_{i\,k} & \ddots & \\ & & & & & & 1 \end{pmatrix} P_k$$



$A = PLU$

Algoritmo:

$[L,U] = PLU \quad (A \in \mathbb{R}^{n\times n} \text{ invertible})$

for $k = 1 : n-1$

  escoger i^* tq $|a_{i^*k}^{(k)}| = \max_{k\le i\le n} |a_{ik}^{(k)}|$

  $F_{A,k}^{(k)} \longleftrightarrow F_{A,i^*}^{(k)}$

  for $i = k+1 : n$

    $l_{ik} = \dfrac{a_{ik}}{a_{kk}}$

    $F_{L,k}^{(k)} \longleftrightarrow F_{L,i^*}^{(k)}$

    for $j = k+1 : n$

      $a_{ij} = a_{ij} - l_{ik} a_{kj}$

    end

  end

end

Ej:

$$\begin{pmatrix} 1 & -6 & 1 \\ 2 & 1 & 0 \\ 4 & 8 & 8 \end{pmatrix}$$

$$A^{(1)} = \begin{pmatrix} 1 & -6 & 1 \\ 2 & 1 & 0 \\ \textcircled{4} & 8 & 8 \end{pmatrix} \Rightarrow \tilde{A}^{(1)} = \begin{pmatrix} & & 1 \\ & 1 & \\ 1 & & \end{pmatrix} \begin{pmatrix} 1 & -6 & 1 \\ 2 & 1 & 0 \\ 4 & 8 & 8 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 8 \\ 2 & 1 & 0 \\ 1 & -6 & 1 \end{pmatrix}$$

$$A^{(2)} = \begin{pmatrix} 4 & 8 & 8 \\ 0 & -3 & -4 \\ 0 & \textcircled{-8} & -1 \end{pmatrix} \qquad L_1 = \begin{pmatrix} 1 & & \\ -1/2 & 1 & \\ -1/4 & & 1 \end{pmatrix}$$

$$\Rightarrow \tilde{A}^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 4 & 8 & 8 \\ 0 & -3 & -4 \\ 0 & -8 & -1 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 8 \\ 0 & -8 & -1 \\ 0 & -3 & -4 \end{pmatrix}$$

$$A^{(3)} = \begin{pmatrix} 4 & 8 & 8 \\ 0 & -8 & -1 \\ 0 & 0 & -29/8 \end{pmatrix} \qquad L_2 = \begin{pmatrix} 1 & & \\ & 1 & \\ & -3/8 & 1 \end{pmatrix}$$

$$A^{(3)} = \begin{pmatrix} 4 & 8 & 8 \\ 0 & -8 & -1 \\ 0 & 0 & -29/8 \end{pmatrix} = U \qquad L_{12} = \begin{pmatrix} 1 & & \\ & 1 & \\ & -3/8 & 1 \end{pmatrix}$$

$$A = P_1 L_1^{-1} P_2 L_2^{-1} U = \begin{pmatrix} & & 1 \\ & 1 & \\ 1 & & \end{pmatrix} \begin{pmatrix} 1 & & \\ 1/2 & 1 & \\ 1/4 & & 1 \end{pmatrix} \begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \begin{pmatrix} 1 & & \\ & 1 & \\ & 3/8 & 1 \end{pmatrix}$$

$$= \tilde{L} U$$

Teo Sea $A \in \mathbb{R}^{n \times n}$ no singular, entonces existe una matriz de permutación $P \in \mathbb{R}^{n \times n}$, $L \in \mathbb{R}^{n \times n}$ (Δ) con los elementos de su diagonal 1's y $U \in \mathbb{R}^{n \times n}$ (∇) tq.

$$PA = LU$$

Estabilidad del Alg. de Sust. regresi-

Análisis regresivo: $\exists\, \check{x}$ tq $f(\check{x}) = \hat{f}(x)$ y

$$\frac{\|x - \check{x}\|}{\|x\|} \leq \eta_f \text{, para } eps \to 0$$

Alg. Sust. Regre. d paso principal es:

$$\forall \quad x_i = \frac{1}{u_{ii}} \left(b_i - \sum_{k=i+1}^{n} x_k u_{ik} \right) \qquad (Ux = b)$$

$$\iff u_{ii} x_i = b_i - \sum_{k=i+1}^{n} x_k u_{ik} \text{, } i = n, \dots, 1$$

Si denotamos por

$$X_{i+1} = [x_{i+1}, \dots, x_n]$$

$n-1$ componentes.

$$U_{i+1} = [u_{i,i+1}, \dots, u_{in}]$$

$n-1$ componentes.

Así, tenemos

$$u_{ii} \cdot x_i = b_i - \langle X_{i+1}, U_{i+1} \rangle \text{, } i = n-1, \dots, 1 \quad \checkmark$$

$$u_{ii} \hat{*} x_i = b_i \hat{-} \langle X_{i+1}, U_{i+1} \rangle_{fl} \text{, } i = n-1, \dots, 1 \quad \checkmark$$

Luego,

$$\langle X_i, U_i \rangle_{fl} = u_{ii} x_i (1 + \varepsilon_{1i})(1 + \varepsilon_{2,i}) + \langle X_{i+1}, U_{i+1} \rangle_{fl} (1 + \varepsilon_{2,i})$$

con $|\varepsilon_{k,i}| \leq eps$, $k = 1, 2$, $i = n-1, \dots, 1$

$$\bar{x} = (x_1, \dots, x_{n-1}) \text{, } \bar{y} = (y_1, \dots, y_{n-1})$$

$$\bar{x} = (x_1, \dots, x_{n-1}), \quad \bar{y} = (y_1, \dots, y_{n-1})$$

$$\langle x, y \rangle = [x_n y_n + \langle \bar{x}, \bar{y} \rangle$$

$$\langle x, y \rangle_{fl} = [x_n \hat{\cdot} y_n \hat{+} \langle \bar{x}, \bar{y} \rangle_{fl}$$

$$= \left[x_n y_n (1+\varepsilon_1) + \langle \bar{x}, \bar{y} \rangle_{fl} \right] (1+\varepsilon_2)$$

Del análisis de estabilidad del p. escalar, sabemos que existe

$$\check{U}_{i+1} = [\check{u}_{i,i+1}, \dots, \check{u}_{i,n}] \quad tq$$

$$\langle x_{i+1}, u_{i+1} \rangle_{fl} = \langle x_{i+1}, \check{U}_{i-1} \rangle \quad y$$

$$\max_{j=i+1,\dots,n} \frac{|u_{ij} - \check{u}_{ij}|}{|u_{ij}|} \doteq (n-i)\, eps \quad (eps \to 0)$$

para $i = n-1, \dots, 1$. Lo cual establece que el Alg. de sust. regresiva es estable en el sentido del A.R por comp.

2021

Nota:

$$L = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & a & 1 & \\ & b & & 1 \end{pmatrix}$$

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$PL = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 0 & 1 \\ 0 & a & 1 & 0 \end{pmatrix}, \quad \boxed{P^{-1} = P^T}$$

$$PLP^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 0 & 1 \\ 0 & a & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 1 & 0 \\ 0 & a & 0 & 1 \end{pmatrix} \qquad PA = LU$$

Alg Doolitle

↳ $LU = A$

↳ Cholesky

$$\begin{pmatrix} a_{11} & a_{1j} & \cdots & a_{1n} \\ \vdots & & & \\ a_{i1} & a_{ij} & \cdots & a_{in} \\ a_{n1} & \cdots & a_{nj} & \cdots a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & & & \\ l_{i1} & l_{ir} & l_{ii} & \\ l_{n1} & l_{ni} & & l_{nn} \end{pmatrix} \cdot \begin{pmatrix} u_{11} & \cdots & u_{1j} & \cdots & u_{1n} \\ & \ddots & & & \\ & & u_{jj} & & a_{jn} \\ & & & & \\ & & & & u_{nn} \end{pmatrix}$$

$$a_{ij} \overset{?}{=} \sum_{r=1} l_{ir} u_{rj}$$

Si $i<j$, estamos por encima de la diagonal de A. Así $l_{ir} = 0$, $r > i$.

Por lo tanto:

$$a_{ij} = \sum_{r=1}^{\min(i,j)} l_{ir} u_{rj}$$

- Tenemos n^2 ecuaciones
- Tenemos $n^2 + n$ incógnitas

Por lo tanto, si fijamos n condiciones tomando $l_{ii} = 1$, $i = 1, \dots, n$. Obtenemos un sistema de ecuaciones que se puede resolver.

$k \le j$

① $$a_{kj} = \sum_{r=1}^{k-1} l_{kr} u_{rj} + \underbrace{l_{kk}}_{1} u_{kj}, \quad j = k, \dots, n$$

Despejamos u_{kj} y tenemos:

$$u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj}, \quad j = k, \dots, n$$

$k < i$

② $$a_{ik} = \sum_{r=1}^{k-1} l_{ir} u_{rk} + l_{ik} u_{kk}, \quad i = k+1, \dots, n$$

Despejamos l_{ik} y tenemos:

$$l_{ik} = \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rk} \right) \frac{1}{u_{kk}}, \quad i = k+1, \dots, n$$

Algoritmo:

[LU] = Doolitle (A)

```
for k = 1 : n
    l_kk = 1
    for j = k : n
        u_kj = a_kj - Σ_{r=1}^{k-1} l_kr u_rj
    end
```

$u_{kj} = a_{kj} - \sum_{r=1}^{k-1} l_{kr} u_{rj}$

end

for $i = k+1 : n$

$l_{ik} = \left(a_{ik} - \sum_{r=1}^{k-1} l_{ir} u_{rk}\right) \frac{1}{u_{kk}}$

end

end

Método de Cholesky

Matrices simétricas definidas positivas.

Def. Una matriz $A \in \mathbb{R}^{n \times n}$ se dice que es:

- semidefinida positiva $x^T A x \geq 0$, $\forall x \in \mathbb{R}^n$
- Definida positiva $x^T A x > 0$, $\forall x \in \mathbb{R}^n \setminus \{0\}$

Si además A es simétrica, se denomina ssdp y sdp respect.

Propiedades: (A es sdp)

1. A es invertible.

   Para $x \neq 0$, $0 < x^T A x \neq 0$

   $Ax = 0$, tiene la solución trivial $x = 0$

   $\ker(A) = 0$.

2. $a_{ii} > 0$, $\forall i = 1, \dots, n$

   $0 < e_i^T A e_i = a_{ii}$

4. Sus valores propios son reales y positivos.

   $Ax = \lambda x$, $x^T A x = \lambda x^T x$

   $\Rightarrow 0 < \frac{x^T A x}{x^T x} = \lambda$

5. Si $B \in \mathbb{R}^{n \times m}$ tq. $\operatorname{rank}(B) = m$. Entonces $B^T \underbrace{A}_{\text{sdp}} B$ es sdp.

6. En cada paso del Alg. E G, cada matriz Principal de A, A_i (se obtiene eliminando la fila i y la columna i de A) es sdp.

7. A es sdp, ssi, $\det(A_k) > 0$, $k = 1, \dots, n$, donde

A_k es la matriz principal de A. (Criterio de Sylvester).

Factorización LDM^T

Es posible encontrar otro tipo de factorización para A removiendo la hipótesis de que los elementos de L son iguales a 1. Específicamente, tendremos fact. de la forma

$$A = LDM^T,$$

donde L, M^T y D son matrices $\triangle$ inf, $\triangledown$ sup y diagonal resp.

En este caso, la resolución de $Ax=b$ se realizará resol.

1. Un sist. $\triangle$ inf $Ly=b$
2. " " diagonal $Dz=y$
3. " " $\triangledown$ sup $M^Tx=z$

$$Ax = b$$

$$LDM^Tx = b$$

Este procedimiento requiere n^2+n flops.

Factorización de Cholesky

martes, 5 de enero de 2021 11:37

$$A = LDM^T$$

En el caso de una matriz simétrica, se tiene $M = L$ y, por lo tanto, $A = LDL^T$. En este caso, el costo computacional se reduce a la mitad.

Teo: (VI) Si todos los menores principales de una matriz $A \in \mathbb{R}^{n\times n}$ son distintos de cero, entonces existe una única matriz diagonal D, una única matriz triang. inf (Δ) L y una única matriz triang. sup (∇) M^T tq. $A = LDM^T$.

Demost. Del resultado de la f. LU sabemos que L y U son únicas si $l_{ii} = 1$, para $i = 1, \dots, n$. Si fijamos los entradas diagonales de D iguales a u_{ii} (que son distintos de cero porque U es no singular), entonces $A = LU = LDD^{-1}U$.

Definiendo $M^T = D^{-1}U$ tenemos la fact. LDM^T, donde $D^{-1}U$ es una matriz (∇). La unicidad de la fact. LDM^T es consecuencia de la unicidad de la F-LU.

Factorización Cholesky para matrices simétricas def.

Factorización Cholesky para matrices simétricas def. positivas.

En este caso, la fact. LDM^T se simplifica considerablemente cuando A es simétrica porque $M = L$. Por otro lado, el costo comp. se reduce a la mitad. con respecto a la fact. LU, ya que es de alrededor de $\frac{n^3}{3}$ flops.

Ej: La matriz de Hilbert de 3×3 admite descomposición LDL^T.

$$H_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{3} & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & \frac{1}{180} \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

En el caso en que A es definida positiva, las entradas diagonales de D son positivas. De esta manera, la idea es proceder como en la eliminación gaussiana y aprovechar la simetría de A.

Si $A \in \mathbb{R}^{n \times n}$ es sdp, tenemos $a^T = [a_{21} \dots a_{n1}]$ y

podemos escribir A de la sig manera:

$$A = \begin{pmatrix} a_{11} & a^T \\ a & \tilde{A} \end{pmatrix}, \quad L_1 = \left(\begin{array}{c|c} 1 & \\ \hline -l & I \end{array}\right)$$

$$l = (l_{21}, \dots, l_{n1})$$

Así, el primer paso de E.G sería:

$$L_1 A = \left(\begin{array}{c|c} a_{11} & a^T \\ \hline 0 & \tilde{A}^{(1)} \end{array}\right)$$

$$L_1 A L_1^T = \left(\begin{array}{c|c} a_{11} & 0 \\ \hline 0 & \tilde{A}^{(1)} \end{array}\right).$$

Así, multiplicamos por las matrices de Frobenius y sus transpuestas por la derecha e izquierda respectivamente. Si continuamos este proceso, obtenemos:

$$L_{n-1} \cdots L_2 L_1 A L_1^T L_2^T \cdots L_{n-1}^T = \begin{pmatrix} a_{11}^{(1)} & & & \\ & a_{22}^{(2)} & & \\ & & \ddots & \\ & & & a_{nn}^{(n-1)} \end{pmatrix}$$

$$= \operatorname{diag}(\underbrace{U}_{F.LU}).$$

Teo (V2 (sdp)): Una matriz $A \in \mathbb{R}^{n \times n}$ es simétrica definida positiva <u>ssi</u> existe una matriz diagonal D

con sus componentes positivas tq.

$\rightarrow A = LDL^T$ (caso particular de la F. LDM^T)

Nota: Como $D = \text{diag}(d_1, \ldots, d_n)$ entonces podemos escribir $D = D^{1/2} D^{1/2}$. En consecuencia, si A es sdp, se tiene que existe $\tilde{L}$ tq.

$$A = \tilde{L} D \tilde{L}^T = \tilde{L} D^{1/2} D^{1/2} \tilde{L}^T$$
$$= LL^T$$

donde $L = \tilde{L} D^{1/2}$.

Resumen:

① $A = LDM^T$ (A partir de la F. LU)

② Si A es sdp $A = LDL^T$

↳ E. G.

↳ $D = D^{1/2} D^{1/2} \Rightarrow A = \tilde{L}\tilde{L}^T$

F. Cholesky.

Para evitarnos cálculos numéricos extras, realizamos esta fact aplicando el Alg. de Doolittle, ya que es más eficiente para matrices sdp al aprovechar la simetría de A.

$$A = \underbrace{\left(\boxed{\cancel{l_{ip}}}\right)}_{L} \underbrace{\left(\boxed{\cancel{l_{kp}}}\right)}_{L^T}$$

$$a_{ik} = \sum_{p=1}^{n} l_{ip} l_{kp} \quad , \quad \text{pero si } i<p \Rightarrow l_{ip} = 0$$

entonces.

$$a_{ik} = \sum_{p=1}^{i} l_{ip} l_{kp} \qquad (1)$$

Si $i > k$, $l_{kp} = 0$

$$a_{ik} = \sum_{p=1}^{k} l_{ip} l_{kp} \qquad (2)$$

De (1) y (2)

$$a_{ik} = \sum_{p=1}^{\min\{i,k\}} l_{ip} l_{kp}$$

Si $i > k$

$$a_{ik} = \sum_{p=1}^{k} l_{ip} l_{kp} = \sum_{p=1}^{k-1} l_{ip} l_{kp} + l_{ik} l_{kk} \qquad (3)$$

si $i = k$

$$\rightarrow \quad a_{kk} = \sum_{p=1}^{k} l_{kp} l_{kp} = \sum_{p=1}^{k-1} l_{kp}^2 + l_{kk}^2 \qquad (4)$$

Despejando l_{kk} de (4) tenemos que:

$$l_{kk} = \left(a_{kk} - \sum_{p=1}^{k-1} l_{kp}^2 \right)^{1/2}$$

y despejando l_{ik} de (3) tenemos que

$$l_{ik} = \left(a_{ik} - \sum_{p=1}^{k-1} l_{ip} l_{kp} \right) / l_{kk}$$

Algoritmo de Cholesky. ($A = LL^T$)

```
[L] = cholesky(A)
for k = 1 : n
    l_kk = (a_kk - sum_{p=1}^{k-1} l_kp^2)^(1/2)
    for i = k+1 : n
        l_ik = (1/l_kk) (a_ik - sum_{p=1}^{k-1} l_ip l_pk)
    end
end
```

Factorización QR

martes, 12 de enero de 2021 11:27

Sea $A \in \mathbb{R}^{n \times n}$, estudiamos la f. QR.

$$A = QR$$

- Q ortonormal
- R $\triangle$ sup.

donde $Q \in \mathbb{R}^{n \times n}$ es ortogonal i.e, $Q Q^T = Q^T Q = I$

Al resolver el sistema $Ax = b$ tendríamos:

$$A = QR$$

$$QRx = b$$

$$Qy = b$$

$$y = Q^T b$$

$$Rx = y \leftarrow \text{Sus Reg.}$$

$$x$$

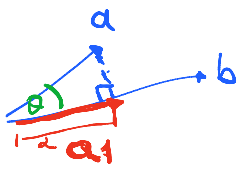
Recordemos de la E-G que en el primer paso tuvemos $A^{(2)} = L_1 A^{(1)}$, así obtuvimos

$$A^{(2)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & & & \\ 0 & & & \\ 0 & & & \end{pmatrix}.$$

La idea en la f. QR es reemplazar las matrices de Frobenius por proyecciones ortogonales sobre el primer eje.

Reflexiones de Householder:

Proyección de un vector sobre otro: La proyección de un vector a sobre un vector b es la proyección ortogonal de a sobre la línea paralela a b.



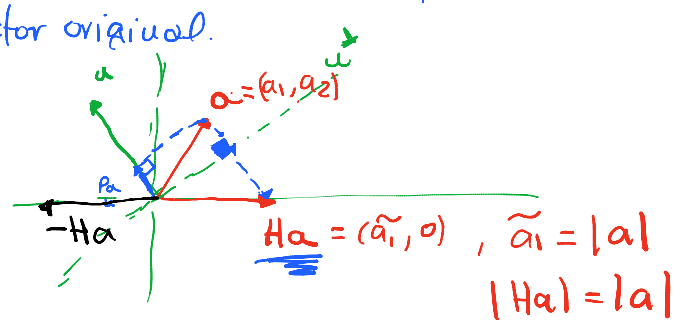
$$a_1 = \alpha \hat{b} \quad , \quad \cos\theta = \frac{a}{|a|} \quad , \quad \cos\theta = \frac{a \cdot b}{|a||b|}$$

donde α es un escalar y $\hat{b}$ es el vector unitario en la dirección de b.

$$\alpha = |a| \cos\theta = a \cdot \frac{b}{|b|} = \frac{\langle a, b \rangle}{|b|}$$

$$P_b^a = a_1 = \langle a, \frac{b}{|b|} \rangle \frac{b}{|b|} = \frac{\langle a, b \rangle}{|b|^2} \cdot b$$

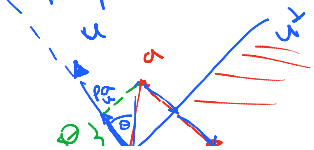
Las reflexiones de Householder son un tipo de transformaciones ortogonales que mantienen la misma del vector original.



$$Ha = (\tilde{a}_1, 0) \quad , \quad \tilde{a}_1 = |a|$$

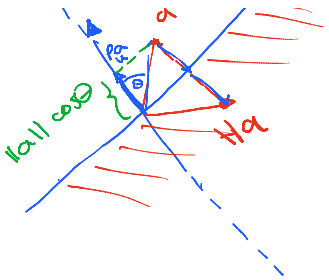
$$|Ha| = |a|$$

Descomponer a en sus componentes que son paralelas y perpendiculares a u.



$$Ha = a - 2 P_u^a$$

$$a - 2 \|a\| \cos\theta \, u$$



$$Ha = a - 2\,p_u^a$$

$$= a - 2\left(\|a\|\cos\theta \cdot \frac{u}{\|u\|}\right)$$

$$= a - 2\,\frac{a^T u}{\|u\|^2}\,u$$

$$= a - 2\,u\,\frac{u^T a}{\|u\|^2}$$

$$= \left(I - \underbrace{\frac{2uu^T}{\|u\|^2}}_{\beta}\right) a$$

H

Propiedades de H (matriz de Householder)

* $H = I - \beta uu^T$, $\beta = \dfrac{2}{\|u\|^2}$

* H es simétrica.

$$H^T = (I - \beta uu^T)^T$$

$$= I - \beta\,(uu^T)^T$$

$$= I - \beta\,uu^T = H$$

* H es ortogonal

$$HH^T = (I - \beta uu^T)(I - \beta uu^T)$$

$$= (I - \beta uu^T)^2$$

$$= I - 2\beta uu^T + \beta^2 (uu^T)(uu^T)$$

$$= I - 2\beta uu^T + \beta^2 u\,(u^T u)\,u^T$$

$$= I - 2\beta uu^T + 2\beta uu^T$$

$= I$.

La utilidad de las transformaciones de Householder radica en que podemos crear ceros en las posiciones $2:n$ de un vector de tamaño n. Ahora, resta determinar el vector u a través del cual se realiza la reflexión.

Tomemos $e_1 = [1, 0, \dots 0]^T$.

* $Ha = (\pm\|a\|_2, 0, \dots, 0)^T$

$$\boxed{Ha = \pm\|a\|_2 e_1}$$

* $H^T H a = a = \pm\|a\|_2 H^T e_1$

$$\Rightarrow H^T e_1 = \pm \frac{a}{\|a\|_2}$$

Ahora $a - Ha$ es paralelo a u y ortogonal a $span(u)^\perp$, entonces podemos tomar

$$u = (a - Ha)$$
$$= a \mp \|a\|_2 e_1$$

De donde se tienen dos opciones

$$\begin{cases} u = a - \|a\|_2 e_1 \\ u = a + \|a\|_2 e_1 \end{cases}$$

vamos a tomar la opción que implique mayor estabilidad.

$$\beta = \frac{2}{\|u\|^2}$$

$$\frac{1}{\beta} = \frac{\|u\|^2}{2} = \frac{(a \mp \|a\|_2 e_1)^T (a \mp \|a\|_2 e_1)}{2}$$

$$= \frac{a^T a \mp 2 a_1 \|a\|_2 + \|a\|_2^2}{2}$$

$$= \frac{\|a\|_2^2 \mp 2 a_1 \|a\| + \|a\|_2^2}{2}$$

$$= \|a\|_2 (\|a\|_2 \mp a_1)$$

Si el ángulo entre a y e_1 es muy pequeño

$\|a\|_2 \approx |a_1|$

$$\frac{1}{\beta} \approx |a_1|(|a_1| \mp a_1).$$

Así, tenemos la resta de dos números muy cercanos que nos pueden conducir a errores de cancelación. Para evitarlo tomamos

$$\frac{1}{\beta} = \|a\| (\|a\| + |a_1|). \text{ y}$$

finalmente tenemos:

$$u = a + \begin{cases} -\|a\| e_1, & a_1 < 0 \\ \|a\| e_1, & a_1 > 0 \end{cases}$$

$$u = a + \text{sign}(a_1)\|a\|e_1$$

Fact. QR

$a_{11}^{(1)} \neq 0$

$$A^{(1)} = A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \\ \vdots & & & \\ a_{n1} & & & a_{nn} \end{pmatrix}$$

$$a^{(1)} = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} \Rightarrow Q_1 = H_1 = I_n - \beta u u^T$$

$$H = I - \beta u u^T = I - \frac{1}{\|a\|(\|a\| + |a_1|)}(a + \text{sign}(a_1)\|a\|e_1)\cdot(a + \text{sign}(a_1)\|a\|e_1)^T$$

$$Ha = \underbrace{-\text{sign}(a_1)\|a\|}_{\sigma} e_1 = \sigma e_1$$

$$A^{(2)} = Q_1(A^1) = \begin{pmatrix} \sigma_1 & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ 0 & a_{22}^{(2)} & & \\ \vdots & \vdots & & \\ 0 & a_{n2}^{(2)} & & a_{nn}^{(2)} \end{pmatrix}$$

$$a^{(2)} = \begin{pmatrix} a_{22}^{(2)} \\ \vdots \\ a_{n2}^{(2)} \end{pmatrix} \Rightarrow Q_2 = \begin{pmatrix} I_1 & \\ & I_{n-1} - \beta u u^T \end{pmatrix}$$

$$A^{(k)} = Q_{k-1} A^{(k-1)} = \begin{pmatrix} \sigma_1 & & a_{1k}^{(2)} & \cdots & a_{1n}^{(2)} \\ & \sigma_2 & & & \\ & & a_{kk}^{(k)} & & a_{kn}^{(k)} \\ & & \vdots & & \vdots \\ & & a_{nk}^{(k)} & & a_{nn}^{(k)} \end{pmatrix}$$

$$Q_k = \begin{pmatrix} I_{k-1} & 0 \\ 0 & I_{n-(k-1)} + \beta u u^T \end{pmatrix}$$

$\Rightarrow$ Así generamos $A^{(1)} \to A^{(2)} \to \cdots \to A^{(n)}$

$$R = A^{(n)} = Q_{n-1} A^{(n-1)} = \cdots = \underbrace{Q_{n-1} Q_{n-2} \cdots Q_1 A^{(1)}}$$

$$= Q^T A$$

$$\Rightarrow R = Q^T A$$

$$\Rightarrow \boxed{A = QR}$$

F. QR

↳ $A = QR$ (Householder)

↳ ortogonal

↳ triang. sup.

Generamos una sucesión de matrices tq.

$$A = A^{(1)} \to A^{(2)} \to A^{(3)} \to \cdots \to A^{(n)}$$

$$R = A^{(n)} = Q_{n-1} A^{(n-1)} = \cdots = \underbrace{Q_{n-1} \cdots Q_1 A^{(1)}}_{Q^T}$$

$$R = Q^T A$$

$$\boxed{QR = A}$$

con

$$Q_k = \begin{pmatrix} I_{k-1} & \\ & I_{n-(k-1)} + \beta u u^T \end{pmatrix}$$

en cada k-ésima iteración.

Entonces, tenemos

$$\underbrace{(I_{n-k+1} - \beta u u^T)}_{2^\circ} \underbrace{A_k}_{1^\circ} = A_k - \beta u (u^T A_k)$$

Así:

```
[QR] = FQR1 (A)
    Q = I
    for k=1:n
        a = [akk, ..., akn]^T   ( A(k:n, k) )
        u = a + sign(a1) ||a||_2 e_1   (norm)
        β = 1 / (||a||_2 (||a||_2 + |a1|))
        Q_k = [ I_{k-1}        0
                0        I_{n-k+1} - β u u^T ]
        A_k = A_k - β u (u^T A_k)
        Q^T = Q Q_k
    end
```

$Ax = b$ $\quad A \in \mathbb{R}^{m \times n}$, $m \geq n$

Lo rango completo por columnas.

QR para matrices rectangulares.

QR para matrices rectangulares.

Una matriz $A \in \mathbb{R}^{m\times n}$ con $m \geq n$ admite fact. QR si existe una matriz ortogonal $Q \in \mathbb{R}^{m\times m}$ y una matriz trapezoidal superior $R \in \mathbb{R}^{m\times n}$ con filas nulas desde la $n+1$ ésima fila $\}$.

$$A = QR$$
$$m\times n \quad m\times m \; m\times n = m\times n$$

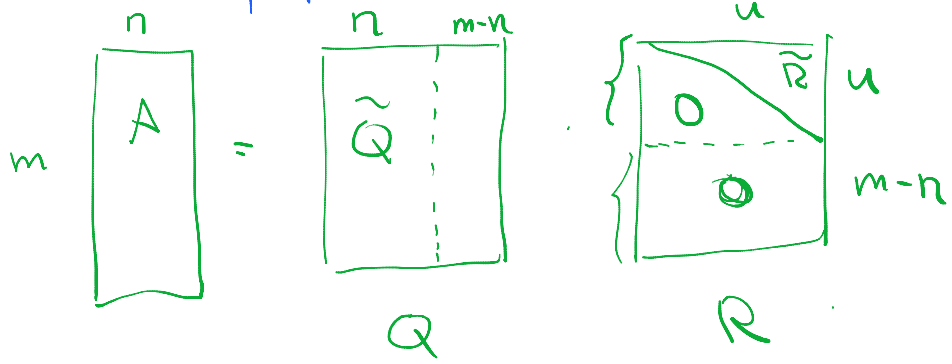
Propiedades: Sea $A \in \mathbb{R}^{m\times n}$ de rango n para la cual existe fact QR conocida. Entonces existe una única fact. de A de la forma

$$A = \tilde{Q}\tilde{R}$$
$$m\times n \quad m\times n \; n\times n = m\times n$$

donde $\tilde{Q}$ y $\tilde{R}$ son submatrices de Q y R dadas respectivamente por

$$\tilde{Q} = Q(1:m, 1:n) \quad \text{y} \quad \tilde{R} = R(1:n, 1:n).$$

Más aún, $\tilde{Q}$ tiene vectores columna ortonormales y $\tilde{R}$ es (∇) triangular superior y coincide con el factor de Cholesky L de la matriz simétrica def. positiva $A^TA = \tilde{R}^T\tilde{R}$.

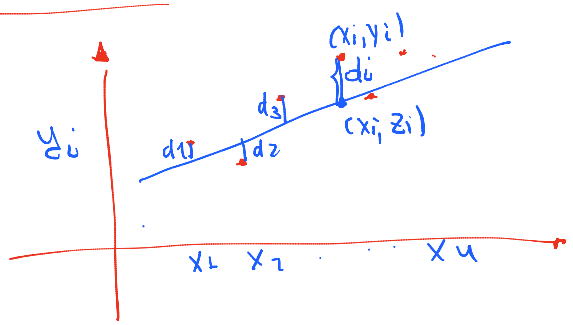


Si A tiene rango completo n, entonces las columnas de $\tilde{Q}$ forman una base ortonormal para el espacio vectorial $\underline{\text{range}(A) = \{Ax,\ x \in \mathbb{R}^n\}}$.

También, si A tiene rango completo, la matriz A^TA ($n\times m \cdot m\times n = n\times n$) es simétrica definida positiva y admite fact. Cholesky única, así

$$L L^T = A^TA = \tilde{R}^T\tilde{Q}^T\tilde{Q}\tilde{R} = \tilde{R}^T\tilde{R}. \quad (\star)$$

Mínimos Cuadrados



$z = ax+b$

Supongamos que tenemos un conjunto de puntos (x_i, y_i) $i = 1, \dots, n$. Se quiere encontrar la recta que pasa más cerca de todos los puntos a la vez. Es decir, se requiere resolver el problema

$$f = \min \sum_{i=1}^{n} d_i^2$$

donde d_i representa la distancia entre el punto (x_i, y_i) dado y el punto de la recta (x_i, z_i). Es decir $d_i = y_i - z_i$, $(z_i = a x_i + b)$,

punto (x_i, y_i) dado y el p...

Es decir $d_i = y_i - z_i$, $(z_i = \alpha x_i + b)$, $i = 1, \dots, n$. Debemos minimizar la función derivable $f(\alpha, b) = \sum_{i=1}^{n} (y_i - z_i)^2$.

$$\rightarrow = \sum_{i=1}^{n} (y_i - \alpha x_i - b)^2,$$

donde α es la pendiente de la recta y b es el punto de corte de la recta en el eje de las ordenadas.

Así, $\min_{\rightarrow (\alpha, b)} f(\alpha, b)$. Este problema tiene condiciones necesarias y suficientes puesto que f es convexa. Las condiciones de optimalidad son:

$$\frac{\partial f}{\partial \alpha} = \sum_{i=1} 2 (y_i - \alpha x_i - b)(-x_i) = 0$$

$$\frac{\partial f}{\partial b} = \sum_{i=1}^{n} 2 (y_i - \alpha x_i - b)(-1) = 0$$

De las dos ecuaciones anteriores se obtiene el sistema cuya solución caracteriza la recta:

$$\underbrace{\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{pmatrix}}_{A^T A \ (2\times 2)} \underbrace{\begin{pmatrix} \alpha \\ b \end{pmatrix}}_{w \ (2\times 1)} = \underbrace{\begin{pmatrix} \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i \end{pmatrix}}_{A^T b \ (2\times 1)}$$

2×2 $\quad 2\times 1$ $\quad 2\times 1$

El sistema anterior se puede escribir como una ecuación normal $\boxed{A^T A w = A^T b}$.

$$A = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \quad w = \begin{pmatrix} \alpha \\ b \end{pmatrix}, \quad b = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Así, tenemos que resolver

$$\boxed{A^T A w = A^T b}.$$

Corrección Examen

(1)

$$B = \begin{pmatrix} 1 & 1-\varepsilon & 3 \\ 2 & 2 & 2 \\ 3 & 6 & 4 \end{pmatrix}$$

$$\left.\begin{array}{l} \det B_1 = 1 \\ \det B_2 = \varepsilon \end{array}\right\} \checkmark$$

$$\det B_3 = \det B_3 = 2\varepsilon + 12$$

- si $\varepsilon = 0$, $\det B_2 = 0 \Rightarrow LU$ no existe
- B es singular si $\varepsilon = -6$, en este caso

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 1.25 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 7 & 3 \\ 0 & -12 & -4 \\ 0 & 0 & 0 \end{pmatrix}$$

Entonces $Ux = y$ tiene infinitas soluciones.

(2) Sea X la inversa de $A \in \mathbb{R}^{n\times n}$ tq. satisface $\boxed{XA = AX = I}$ Así, para $j = 1, \dots, n$ tenemos $\underline{Ax_j = e_j}$. Después de hacer la f. LU de A, la inversa se obtiene al resolver n sistemas lineales con la misma A y diferente lado derecho.

(3) Supongamos que $A = LL^T$. Debemos demostrar $x^TAx \geq 0$.

Sea $A = LL^T$,

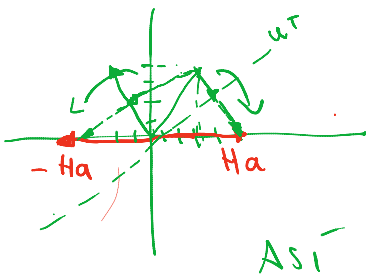
$$x^TAx = x^T(LL^T)x = \langle L^Tx, L^Tx\rangle$$
$$= \|L^Tx\|^2 \geq 0 \quad \blacksquare$$

(4) $H^2 = I$

$$H^2 = (I - \beta uu^T)^2$$
$$= I - 2\beta uu^T + \beta^2(uu^T)(uu^T)$$
$$= I - 2\beta uu^T + \beta^2 u(u^Tu)u^T$$
$$= I - 2\beta uu^T + \frac{4}{(u^Tu)^2}u(u^Tu)u^T$$
$$= I - 2\beta uu^T + \frac{2\cdot 2}{u^Tu}uu^T$$
$$= I - 2\beta uu^T + 2\beta uu^T$$
$$= I.$$

(5) $H = (\pm\|a\|, 0)$, $\|a\| = 5$

(5)



$H = (\pm\|a\|, 0)$, $\|a\| = 5$

$u = a + \|a\| e_1$

$u = a - \|a\| e_1$

Así $u = a - \|a\| e_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix} - 5 \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$= \begin{pmatrix} 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 5 \\ 0 \end{pmatrix}$

$= \begin{pmatrix} -2 \\ 4 \end{pmatrix}$

$Ha = \begin{pmatrix} 5 \\ 0 \end{pmatrix}$

$A^TA = \tilde{A} = LL^T$

$u \times u$ $\quad u \times n \times n \times u$

(6) Sea $A \in \mathbb{R}^{m \times n}$, $A^TA = LL^T$

$n \times m \; m \times n$

y $Q = A(L^T)^{-1}$. Prueba que Q es ortogonal

$QQ^T = I$, $A^TQ = L$

$A^TA = \underbrace{A^TQ}_{L} \underbrace{(A^TQ)^T}_{L^T}$

$= A^TQ\,Q^TA$

$n \times m \quad m \times n$

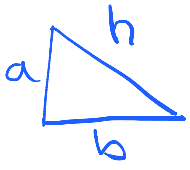
Esto es posible sii $QQ^T = I$.

QR y Cholesky.

$A^TA = R^TQ^T(QR) = R^TR = LL^T$, $R^T = L$.

Prueba

Considere el problema de calcular el cateto b de un triángulo rectángulo a partir de su hipotenusa a y el cateto a. Es decir, se quiere analizar la función f tq. $(a,h) \to b$. Asumiendo que la operación raíz cuadrada tiene un indicador de estabilidad σ_r y un número de condición K_r, hallar una estimación para $\sigma_f K_f$.



$$f(a,h) = b = \sqrt{h^2 - a^2}$$

$$g(z) = \sqrt{z}$$

$$y(a,h) = h^2 - a^2$$

$$f(a,h) = g(y(a,h))$$

Sin embargo $y(a,h) = l(p(a,h))$, $p(a,h) = (a^2, h^2)$

$= l(a^2, h^2)$ $= (a \cdot a, h \cdot h)$

$= h^2 - a^2$

$$\sigma_f K_f \leq \sigma_g K_g + \sigma_y K_y K_g \quad , \; \sigma_g := \sigma_r \; , \; K_g := K_r$$

$$= \sigma_r K_r + \sigma_y K_y K_r$$

$$\sigma_y K_y \leq \underbrace{\sigma_l K_l}_{\leq 1} + \underbrace{\sigma_p K_p K_l}_{\leq 1}$$

$$= 1 + k\ell$$

$$= 1 + \frac{|a^2| + |h^2|}{|a^2 - h^2|}$$

$$\Rightarrow G_{fkf} = \sigma_g k_g + \left(1 + \frac{|a^2| + |h^2|}{|a^2 - h^2|}\right) k_g$$

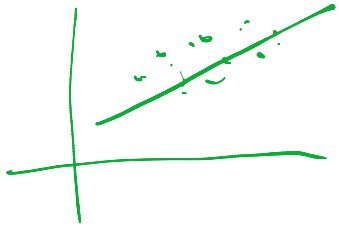
Mínimos Cuadrados

martes, 26 de enero de 2021 12:00

Resumen.

→ $A \in \mathbb{R}^{m \times n}$, existe la fact. QR.

$A = QR$

→ 

$\Rightarrow$ $\overbrace{A^T A x = A^T b}^{2 \times 2}$

$x = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ α = pendiente, β = corte

→ Sea $A \in \mathbb{R}^{m \times n}$ y $b \in \mathbb{R}^m$, encont

$x \in \mathbb{R}^n$ tq

$Ax = b$ → no necesariamente la solución es única o existe.

$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 = \min_{x \in \mathbb{R}^n} f(x)$

i.e $\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 = \min_{x} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} x_j - b_i \right)^2$

Ejemplo: $A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 2 \end{bmatrix}$, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$b = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$

$$f(x) = \|Ax - b\|^2$$

$$= (2x_1 - 1)^2 + (-x_1 + x_2)^2 + (2x_2 + 1)^2$$

Derivando $f(x)$ tenemos:

$$\nabla f(x) = \begin{pmatrix} 4 \cdot (2x_1 - 1) - 2(-x_1 + x_2) \\ 2 \cdot (-x_1 + x_2) + 4(2x_2 + 1) \end{pmatrix}$$

$$\nabla f(x) = \begin{pmatrix} 10x_1 - 2x_2 - 4 \\ -2x_1 + 10x_2 + 4 \end{pmatrix}$$

Así $\boxed{\nabla f(\hat{x}) = 0}$, obtenemos $\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{1}{3} \end{pmatrix}$

* $\hat{r} = A\hat{x} - b$ es el vector residuo

* Si $\hat{r} = 0$, entonces $\hat{x}$ es solución de $Ax = b$

* Si $\hat{r} \neq 0$, entonces $\hat{x}$ es una solución aproxi-

* Si $\hat{r} \neq 0$, entonces $\hat{x}$ es una solución aproximada de $Ax=b$.

* Si A tiene columnas l.i, entonces A es de rango completo por columnas, y A^TA es invertible y

$$\boxed{\hat{x} = (A^TA)^{-1}A^Tb}$$

es solución única del problema.

$$\boxed{\min_{x\in\mathbb{R}} \|Ax-b\|^2} \iff \boxed{A^TAx = A^Tb}$$

cod op.

Si A es rango comp. por colu. la solución es única

Teo: Sea V un espacio vectorial Euclideano de dimensión finita con producto escalar $\langle\cdot,\cdot\rangle$, $U \subset V$ un subespacio y $U^\perp = \{v \in V \mid \langle v,u\rangle = 0 \ \forall u \in U\}$ es el complemento ortogonal en V. Para todo $v \in V$ tenemos la sig. propiedad respecto a la norma inducida por el producto escalar ($\|v\| = \sqrt{\langle v,v\rangle}$):

$\|v\| = \sqrt{\langle v, v \rangle}$

$\|v - u\| = \min_{u' \in U} \|v - u'\| \iff v - u \in U^{\perp}$

Demost.- Sea $u \in U$ un punto únicamente determinado tq. $v - u \in U^{\perp}$. Luego, para todo $u' \in U$ tenemos:

$$\|v - u'\|^2 = \|v - u + u - u'\|^2 =$$

$$= \|v - u\|^2 + 2\langle v - u, u - u' \rangle + \|u - u'\|^2$$

($\langle v - u, u - u' \rangle = 0$)

$$= \|v - u\|^2 + \|u - u'\|^2$$

$$\geq \|v - u\|^2$$

$$\iff \|v - u'\|^2 \geq \|v - u\|^2 \quad \forall u' \in U$$

$\forall u' \in U$

Teo: El vector $x \in \mathbb{R}^n$ es solución del problema de mínimos cuadrados

$$\min_x \|b - Ax'\| = \|b - Ax\| \quad \text{ssi} \quad x$$

satisface la ecuación normal

$$A^TAx = A^Tb$$

En particular, el problema de mínimos cuadrados tiene solución única ssi el rango de A es máximo por columnas.

Demost. Aplicamos el Teo anterior a $V = \mathbb{R}^m$ y $U = R(A)$, así

$$\|b - Ax\| = \min_{x' \in \mathbb{R}^n} \|b - Ax'\|$$

$$\Leftrightarrow \langle b - Ax, Ax' \rangle = 0, \quad \forall x' \in \mathbb{R}^n$$

$$\Leftrightarrow \langle A^T(b - Ax), x' \rangle = 0, \quad \forall x' \in \mathbb{R}^n$$

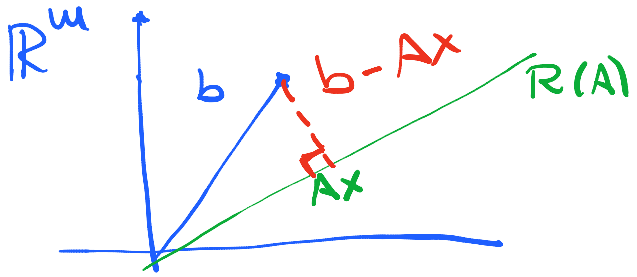
$$\Leftrightarrow A^Tb - A^TAx = 0$$

$$\Leftrightarrow A^TAx = A^Tb.$$

La unicidad de la solución x se da gracias a que A^TA es invertible ssi rango$(A) = n$, i.e., (A es rango comp. por columnas).

columnas).

Nota: Geométricamente, la ecuación normal se determina gracias a que el vector $b - Ax$ es normal a $\mathcal{R}(A) \subseteq \mathbb{R}^n$.



En general si un sistema está sobrecondicionado y no es posible hallar x tq. $Ax = b$, al menos podemos buscar x tq. $\|Ax-b\|^2$ sea mínima.

Resultado Alternativo:

Sea $f(x) = \|Ax-b\|^2 = \langle Ax-b, Ax-b \rangle$

$$= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} x_j - b_i \right)$$

Encontremos las derivadas parciales de f respecto de x_k.

$$\frac{\partial f}{\partial x_k}(x) = 2\left(\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}x_j - b_i\right)(A_{ik})$$

$$= 2\sum_{i=1}^{m} A_{ik}\sum_{j=1}^{n} A_{ij}x_j - b_i$$

$$= 2\left(A^T(Ax-b)\right)_k$$

El gradiente de f es

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right)$$

$$= 2A^T(Ax-b)$$

Si $\hat{x}$ minimiza $f(x)$ entonces satisface $\nabla f(\hat{x}) = 2A^T(A\hat{x}-b) = 0$

$$\Rightarrow A^TA\hat{x} - A^Tb = 0$$

$$A^T A \hat{x} = A^T b$$

si $A^T A$ es invert. → Cholesky LL^T

↳ QR

$f: \mathbb{R}^n \to \mathbb{R}$

$\nabla f: A^T A x - A^T b \in \mathbb{R}^n$

$n \times n \times n \times 1$ → $n \times 1$

$n \times n \times n \times 1$ → $n \times 1$

Métodos Iterativos

miércoles, 27 de enero de 2021 12:06

$Ax = b$

↳ Métodos directos

↳ $LU = A$

↳ cholesky, $A = LL^T$

↳ $QR \Rightarrow A = QR$

↳ Métodos iterativos.

$x^{(0)} \rightarrow x^{(1)} \rightarrow x^{(2)} \rightarrow \cdots \rightarrow x^*$

- Métodos directos comienzan con matrices arbitrariamente llenas o densas.
- El costo de resolver sistemas es del orden $O(n^3)$
- Sin embargo, existen sistemas $Ax = b$ donde la matriz es estructurada y la mayoría de componentes son cero.

La idea de los métodos iterativos es usarlos para sistemas lineales donde las matrices tienen dimensiones grandes y matrices con cierta estructura donde la mayoría de sus entradas son cero.

En este caso, los método iterativos generan una sucesión de soluciones a través de iteraciones de punto fijo:

Nota Ecuación de punto fijo

Nota Ecuación de punto fijo

$$\phi(x) = x$$

Construimos una sucesión $\{x_0, x_1, \dots\}$ con la ayuda de un esquema iterativo:

$$x_{k+1} = \phi(x_k) \quad \text{con } k = 0, 1, \dots$$

Se espera que la sucesión $\{x_k\}$ converja a un punto fijo x^* tq.

$$\phi(x^*) = x^*$$

En general, un método iterativo para sist. lineales tiene la forma

$$x^{k+1} = \phi(x^1, x^2, \dots, x^k).$$

Es decir, $k+1$-ésimo valor de la sucesión está en función de los valores de las iteraciones anteriores por medio de la función ϕ.

Sea $A \in \mathbb{R}^{n \times n}$ una matriz no singular y $b \in \mathbb{R}^n$, buscamos una solución para el sistema $Ax = b$ usando una sucesión de aproximaciones.

$Ax = b$, $\phi(x) = x$

La idea es descomponer A de la siguiente manera: $\boxed{A = M - N}$

considero:

$$Mx - Nx = b$$

$$Mx = Nx + b$$

Si M es invertible entonces

$$x = M^{-1}Nx + M^{-1}b$$

$$= Bx + c$$

De esta manera generamos un método iterativo de la forma:

$$\boxed{x^{k+1} = M^{-1}Nx^k + M^{-1}b}$$

$$= Bx^k + c = \phi(x^k)$$

$B = M^{-1}N$, $c = M^{-1}b$.

Así tenemos el siguiente esquema iterativo

$$(*) \quad x^{k+1} = Bx^k + c, \quad k = 1, 2, \ldots$$

Los métodos expresados por $(*)$ se denominan métodos iterativos estacionarios debido a que la matriz B, llamada matriz de iteración del método es fija.

La construcción del método obliga

a que, si la sucesión generada por (x) converge a un punto fijo de la función $x \mapsto Bx+c$, en consecuencia también es solución del sist. lineal

$$Ax=b.$$

La descomposición $A=M-N$ se la puede realiza de muchas maneras. Sin embargo, es necesario seleccionar una matriz M que sea fácil de invertir.

Métodos iterativo Clásicos.

$A \in \mathbb{R}^{n\times n}$. $A=L+D+U$

$$L=\begin{bmatrix} 0 & & & 0 \\ a_{21} & 0 & & \\ \vdots & & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix}$$

$$D=\begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix}$$

$$U=\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & 0 & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{bmatrix}$$

$$A=L+D+U$$

① Método de Richardson.

$$A=M-N,$$

$$M=I \qquad B=M^{-1}N$$

$M = I$, $B = M^{-1}N$

$N = I - A$,

así, el esquema iterativo es el siguiente:

$$x^{k+1} = Bx^k + c$$
$$= (I - A)x^k + c$$
$$= x^k - Ax^k + b$$

(2) Método de Jacobi , $A = M - N$, $A = L + D + U$

$M = D$

$N = -(L+U)$

$\Rightarrow B = (I - D^{-1}A)$

$B = -D^{-1}(L+U)$

Así, $x^{k+1} = -D^{-1}(L+U)x^k + D^{-1}b$

(3) Método de Gauss-Seidel

$M = L + D$, $N = -U$

$B = -(D+L)^{-1}U$

$B = (I - (D+L)^{-1}A)$

$$x^{k+1} = (I - (D+L)^{-1}A)x^k + (D+L)^{-1}b$$

Resumen $x^{k+1} = \phi(x^k)$, $k = 0, 1, \ldots$

$Ax = b$, $A = M - N$

$\phi x^k = M^{-1}Nx^k + M^{-1}b = Bx^k + c$

- Richardson (B)
- Jacobi (B)
- Gauss-Seidel (B)

$$x^{k+1} = Bx^k + c \qquad (**)$$

Convergencia de los métodos it. estacionarios.

Supongamos que c es cero, $c = 0$. La iteración (**) esta relacionada con la ec. de punto fijo $x = Bx$. Este problema corresponde a hallar un vector propio x asociado al valor propio 1. Es así que un método iterativo se puede analizar basándose en las propiedades espectrales de la matriz de iteración B.

Recordemos que el radio espectral se define como

$$\rho(A) = \max_{1 \le i \le n} |\lambda_i(A)|$$

donde $\lambda_i(A) \in \mathbb{C}$, $i = 1, \dots, n$ representa los valores propios de A.

Lema: Sea $B \in \mathbb{R}^{n \times n}$, entonces la equivalencia

$$\lim_{k \to \infty} B^k = 0 \iff \rho(B) < 1$$

es verdadera.

Demost: $\Rightarrow$ Podemos $\lim B^k = 0$ implica que

$$\lim_{k\to\infty} B^k x = 0 \quad \forall x \in \mathbb{R}^n$$

Tomemos $x \neq 0$.

$$0 \leq \frac{1}{\|x\|} \left\| \lim_{k\to\infty} B^k x \right\| \leq \frac{1}{\|x\|} \lim_{k\to\infty} \|B^k\| \|x\|$$

$$= \lim_{k\to\infty} \|B^k\| = 0$$

$$0 \leq \left\| \lim_{k\to\infty} B^k x \right\| \leq 0$$

$$\Rightarrow \lim_{k\to\infty} B^k x = 0 \quad \forall x \in \mathbb{R}^n$$

En particular si x un vector propio unitario de B con valor propio asociado λ.

$$0 = \lim_{k\to\infty} B^k x = \lim_{k\to\infty} \lambda^k x$$

$$\Rightarrow 0 = \left\| \lim_{k\to\infty} \lambda^k x \right\| = \lim_{k\to\infty} \|\lambda^k x\|$$

$$= \lim_{k\to\infty} |\lambda^k| \cancel{\|x\|}^{1}$$

$$\Rightarrow 0 = \lim_{k\to\infty} |\lambda^k|$$

esto último se da sólo si $|\lambda| < 1$. En consecuencia $\rho(B) < 1$.

$\Leftarrow$ El recíproco es cierto utilizando la siguiente propiedad (Quarteroni Prop. 1.13): Para $\varepsilon > 0$ existe al menos una norma matricial consistente $\|\cdot\|$, la cual depende de la matriz B tq.

$$\|B\| < \rho(B) + \varepsilon.$$

El recíproco se sigue del resultado anterior.

Con este resultado podemos establecer el

El teorema se sigue del resultado anterior.

Con este resultado podemos establecer el resultado de convergencia general para M. it. estacionarios.

Teo El método iterativo estacionario de la forma $x^{k+1} = Bx^k + c$ es convergente para cualquier valor inicial $x^{(0)}$ **SSI** $\rho(B) < 1$.

Demos. Si restamos la ecuación de punto fijo $x = Bx + c$ de $x^{k+1} = Bx^k + c$

se obtiene:

$$\begin{aligned} x^{k+1} &= Bx^k + c \\ - \quad x &= Bx + c \end{aligned}$$

$$\Rightarrow \quad x^{k+1} - x = B(x^k - x).$$

De donde, podemos generar una relación de recurrencia para el error absoluto, i.e.,

$$\begin{aligned} x^k - x &= B(x^{k-1} - x) \\ &= B(B(x^{k-2} - x)) \\ &= B^2(x^{k-2} - x) \\ &\vdots \\ &= B^k(x^0 - x). \end{aligned}$$

Ahora, tomando el límite en ambos lados

$$\lim_{k \to \infty} \|x^k - x\| = \lim_{k \to \infty} \|B^k(x^0 - x)\| = 0$$

y gracias al resultado anterior y sup. que $\rho(B) < 1$, tenemos que

y gracias al resultado anterior y sup. que $\rho(B) < 1$, tenemos que

$$X^{(k)} \to X \quad \text{si } k \to \infty. \quad \blacksquare$$

Nota. Una implicación del Teo * es que el método iterativo estacionario es convergente para cualquier valor inicial $x^{(0)}$ si y solamente si (ssi) $\|B\| < 1$, ya que para una matriz $A \in \mathbb{R}^{n \times n}$ se sigue que $\rho(A) \leq \|A\|$, donde $\|A\| = \sup_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|}$ para alguna norma vectorial $\|\cdot\|$. De la misma manera, se sigue del último teo. que un método it. estacionario converge para cualquier valor inicial $x^{(0)}$ si para alguna norma matricial se tiene que $\|B\| < 1$.

Convergencia de Métodos it. clásicos.

Convergencia del método de Richardson:-

En este caso, la matriz de iteración $B = I - A$. Consideremos una generalización y tomemos $B = I - \omega A$, $\omega > 0$, fijo.

Calculemos los valores propios de B. Si λ_i, $i = 1, \dots, n$ son los valores propios de A sus respectivos vectores propios asociados son x_i, $i = 1, \dots, n$.

Entonces:

$$B x_i = (I - \omega A) x_i$$

$$= x_i - \omega A x_i$$
$$= x_i - \omega \lambda_i x_i$$
$$= (1 - \omega\lambda_i)\, x_i$$
$$\mu_i$$

es decir, los valores propios de $B = I - \omega A$ corresponden a los $\mu_i := 1 - \omega\lambda_i$, $i = 1, \dots, n$.

Ahora,

$$1 - \omega \max_i \lambda_i \leq \mu_i = 1 - \omega\lambda_i \leq 1 - \omega \min_i \lambda_i$$

Así, suponemos $0 < \min_i \lambda_i$ y $0 < \omega < \frac{2}{\max \lambda_i}$

de donde,

$$-1 = 1 - \frac{2\max \lambda}{\max \lambda_i} < 1 - \omega \max_i \lambda_i$$
$$\leq \mu_i$$
$$\leq 1 - \omega \min_i \lambda$$
$$< 1$$

$$\Rightarrow \quad -1 \leq \mu_i \leq 1 \quad \Rightarrow \quad |\mu_i| < 1.$$

Por lo tanto, el método de Richardson converge para cualquier v inicial y para $B = I - \omega A$, siempre que $\min_i(\lambda_i) > 0$ y si $0 < \omega < \frac{2}{\max_i (\lambda_i)}$.

★ $\rho(B) < 1 \iff$ M. Iterativo estacionario converge.

ssi

* $|\rho(B)| < 1 \iff$ SSi el método iterativo estacionario converge.

→ Richardson: si $\begin{cases} 0 < \omega < \dfrac{2}{\max_i(\lambda_i(A))} \\ 0 < \min_i \lambda_i(A) \end{cases} \Rightarrow$ Richardson converge

→ Jacobi: Si Matriz A es diagonal dominante $\Rightarrow$ entonces Jacobi converge.

→ Gauss-Seidel: si A es s.d.p. $\Rightarrow$ entonces G-S converge.

Convergencia del Método de Jacobi

Recordemos que $B = D^{-1}(L+U)$

$= -D^{-1}L - D^{-1}U$,

de esta manera analizamos las entradas b_{ij} de B.

$D = \begin{pmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{pmatrix}$, $D^{-1} = \begin{pmatrix} \frac{1}{a_{11}} & & & \\ & \frac{1}{a_{12}} & & \\ & & \ddots & \\ & & & \frac{1}{a_{nn}} \end{pmatrix}$

$L = \begin{pmatrix} 0 & & & \\ a_{21} & 0 & & \\ a_{31} & & \ddots & \\ \vdots & & & 0 \\ a_{n1} & & & \end{pmatrix}$, $U = \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}$

$$b_{ij} = \begin{cases} \dfrac{a_{ij}}{a_{ii}} & i \neq j \\ 0 \text{ si } i = j \end{cases}$$.

De donde si A es diag. dominante tenemos

$$|a_{ii}| > \sum_{j \neq i}^{n} |a_{ij}| \quad , \forall i = 1, \dots, n$$

$$\Rightarrow \quad 1 > \sum_{j=1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \quad , \forall \ i = 1, \dots n$$

$$\Rightarrow \quad 1 > \sum_{i}^{n} |b_{ii}| \quad \forall i = 1, \dots, n$$

$$\Rightarrow \quad 1 > \sum_{j=1}^{n} |b_{ij}| \quad , \quad \forall i = 1, \dots, n$$

Recordemos la norma infinito.

$$\|B\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |b_{ij}| < 1$$

$\Rightarrow \rho(B) < 1$. Así, el método de Jacobi es convergente para cualquier valor inicial si A es diagonal dominante por filas o columnas.

Convergencia del Método de Gauss-Seidel.

En este caso $B = I - (D+L)^{-1}A$. Gauss-Seidel es adecuado para matrices simétricas definidas positivas. Requerimos introducir el siguiente producto escalar en $\mathbb{R}^4$:

$$(\cdot,\cdot)_A : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$$

asociado a la matriz A de la sig. manera

$$(x, y)_A = y^T A x.$$

con A simétrica def. positiva.

Por otro lado, introducimos la matriz adjunta de B (matriz de iteración) con respecto a este producto escalar.

Nota. Sea V un espacio vectorial donde se definen un op. lineal T y un producto escalar $\langle \bar{u}, \bar{v} \rangle$. Entonces se cumple:

$$\langle T\bar{u}, \bar{v} \rangle = \langle \bar{u}, T^* \bar{v} \rangle$$

$$\langle T\bar{u}, \bar{v} \rangle = \langle \bar{u}, T^* \bar{v} \rangle$$

El operador adjunto está dado por T^*.

$$(Bx, y)_A = y^T ABx = y^T AB A^{-1} Ax$$

$$= ((ABA^{-1})^T y)^T Ax$$

$$= (x, (ABA^{-1})^T y)_A$$

$$= (x, B^* y)_A$$

Así, $B^* = (ABA^{-1})^T = A^{-1} B^T A$.

Antes de probar que G-S converge para matrices s.d.p, introducimos el siguiente Teo.

Teo. Sea B y su adjunta B^* con respecto al producto escalar $(\cdot,\cdot)_A$. Si $\tilde{B} = I - B^* B$ satisface

$$(\tilde{B}x, x)_A > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\},$$

entonces $\rho(B) < 1$.

Demostración: Tenemos que para $x \neq 0$:

$$0 < (\tilde{B}x, x)_A = ((I - B^* B)x, x)_A \quad \text{Deber}$$

$$= (x, x)_A - (B^* Bx, x)_A$$

$$= \|x\|_A^2 - (Bx, Bx)$$

$$= \|x\|_A^2 - \|Bx\|_A^2 .$$

$$\Rightarrow \|Bx\|_A^2 < \|x\|_A^2 .$$

$$\Rightarrow \|Bx\|_A^2 < \|x\|_A^2$$

$$\frac{\|Bx\|_A^2}{\|x\|_A^2} < 1$$

$$\sup_{x \neq 0} \frac{\|Bx\|_A}{\|x\|_A} < 1$$

$$\Rightarrow \|B\| < 1 \Rightarrow \rho(B) < 1$$

Ahora, debemos verificar que la matriz de iteración B del método de G-S verifica que $\tilde{B} = I - B^* B$ es definida positiva con respecto al product. esc. $(\cdot,\cdot)_A$.

$$B^* = A^{-1} B^T A$$

$$= A^{-1} (I - (D+L)^{-1} A)^T A$$

$$= A^{-1} (I - A^T ((D+L)^{-1})^T) A$$

$$= A^{-1} A - ((D+L)^{-1})^T A \quad (\text{pues } A \text{ es simétrica})$$

$$= I - ((D+L)^T)^{-1} A$$

$$= I - (D+U)^{-1} A$$

Ahora, calculemos $\tilde{B} = I - B^* B$.

$$\tilde{B} = I - B^* B$$

$$= I - (I - (D+U)^{-1} A)(I - (D+L)^{-1} A)$$

$$= I - I + (D+L)^{-1} A + (D+U)^{-1} A - (D+U)^{-1} A (D+L)^{-1} A$$

$$= \left[(D+L)^{-1} + (D+U)^{-1} - (D+U)^{-1} A (D+L)^{-1} \right] A$$

$$= (D+U)^{-1} \left[(D+U)(D+L)^{-1} + I - A(D+L)^{-1} \right] A$$

$$= (D+U)^{-1} \left[(D+U) + (D+L) - A \right] (D+L)^{-1} A$$

$$= (D+U)^{-1} \left[D + U + D + L - (L + D + U) \right] (D+L)^{-1} A$$

$$= (D+U)^{-1} [D] (D+L)^{-1} A$$

$$= (D+U)^{-1} D (D+L)^{-1} A$$

Verifiquemos que $\tilde{B}$ es def. posit. con respecto al $(\cdot,\cdot)_A$, para $x \neq 0$,

$$(\tilde{B}x, x)_A = \left(\left[(D+U)^{-1} D (D+L)^{-1} A \right] x, x \right)_A$$

$$= x^T A \left[(D+U)^{-1} D (D+L)^{-1} A \right] x$$

$$= \left((D+U)^{-1^T} A x \right)^T D (D+L)^{-1} A x$$

$$= \left(\left(D + L^T \right)^{-1} \right)^T A x \Big)^T \quad " \quad "$$

$$= \left(\left((D+L^T)^T \right)^{-1} A x \right)^T \quad " \quad "$$

$$= \left((D+L)^{-1} A x \right)^T D \, (D+L)^{-1} A x \quad \forall x \neq 0$$

$$= \xi^T D \xi > 0 \quad \forall \xi \neq 0$$

$$= \langle \xi, D\xi \rangle > 0 \text{ pues } D \text{ es s.d.p}$$

$$\Rightarrow (\tilde{B}x, x) > 0 \quad \forall x \neq 0.$$

Del Teo anterior, concluimos que $\rho(B) < 1$. y por lo tanto, el método de G-S es convergente para matrices s.d.p.

Tasa y velocidad de convergencia:

Definición. Sea una sucesión $(x^{(k)})_{k \in \mathbb{N}}$ tq. $x^{(k)} \to x$ cuando $k \to \infty$. Se dice que la sucesión $(x^{(k)})_{k \in \mathbb{N}}$ converge a x con orden r ssi:

$$\lim_{k \to \infty} \frac{\|x^{(k+1)} - x\|}{\|x^{(k)} - x\|^r} = C$$

donde C es una constante.

- Si $r = 1$ y $C < 1$, la convergencia es lineal.
- Si $r > 1$, la convergencia es superlineal.
- Si $r = 2$, la $\|\ \|$ es cuadrática.



Recordemos:

$$b_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}} & \text{si } i \neq j \\ 0 & \text{si } i = j \end{cases}$$

Jacobi:

$$x^{k+1} = \underbrace{-D^{-1}(L+U)}_{B} x^k + D^{-1} b$$

$x = -D^{-1}(L+U)x$ ...

$$x^{k+1} = \begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \\ \vdots \\ x_i^{k+1} \\ \vdots \\ x_n^{k+1} \end{pmatrix}$$

Construimos la solución en la it $k+1$-ésima componente a componente.

$$x_i^{k+1} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right], \quad i = 1, \ldots, n$$

```
[x, iter, err] = Jacobi1 (A, b, x0, tol)
n = size(A,1)
iter = 0
r = b - A*x0
r0 = norm(r)
err = norm(r)
x = x0
while err > tol
    for i = 1:n
        s = 0
        for j = 1 : i-1
            s = s + A(i,j) * x(j)
        end for
        for j = i+1 : n
            s = s + A(i,j) * x(j)
        end
        x(i) = (b(i) - s) / A(i,i)
    end
```

```
    end
    r = b - Ax
    err = norm(r) / r0
    iter = iter + 1
end
```

Gauss-Seidel:

$$x^{k+1} = (I - (D+L)^{-1}A)\,x^{k} + (D+L)^{-1}b$$

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}\right]$$

para $i = 1, \ldots, n$.

$$x^{k+1} = \begin{pmatrix} x_1^{k+1} \\ \vdots \\ x_i^{k+1} \\ \vdots \\ x_n^{k+1} \end{pmatrix}$$

Nota: guardar el vector x en la iteración anterior.

Matriz de prueba:

```
n =
e = ones(n,1)
A = spdiags([e 2e e], -1:1, n, n);
full(A)
```

$\hat{x} = A \backslash b$

↓ Matlab

JOR - SOR

Resumen: En el método de Jacobi, una vez elegido un vector inicial x^0, x^{k+1} se calcula con la fórmula:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right], \quad i = 1, \ldots, n$$

que equivale a realizar la iteración

$$a) \quad x^{k+1} = D^{-1}(L+U)x^k + D^{-1}b.$$

La generalización del M. de Jacobi es la sobre-relajación del método o JOR (Jacobi over relaxation method), en el cual, se introduce el parámetro de relajación ω y la iteración a) es reemplazada por

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right] + (1-\omega) x_i^{(k)}, \quad i = 1, \ldots, 4$$

El método funciona para $\omega \neq 0$, con $\omega = 1$ coincide con el método de Jacobi.

JOR

A, b, x_0, nmax, tol, ω

iter = 0

$r = b - Ax_0$

$r_0 = \text{norm}(r)$

```
err = norm(r)
x = x0
while err > tol and iter < nmax
    for i = 1:n
        s = 0
        for j = 1 : i-1
            s = s + A(i,j)*x(j)
        end
        for j = i+1 : n
            s = s + A(i,j)*x(j)
        end
        x(i) = w*(b(i) - s)/A(i,i) + (1-w)*x(i)
    end
    r = b - Ax
    err = norm(r)/r0
end
```

En el método de Gauss-Seidel difiere del método de Jacobi en el hecho de que en la it. $k+1$ los valores $x_i^{(k+1)}$ disponibles son usados para actualizar la información, i.e., tenemos:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left[b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}\right], \quad i = 1, \ldots, n \qquad (b)$$

$, i = 1, \ldots, n$ (b)

que equivale a realizar

$$x^{(k+1)} = (I - (D+L)^{-1} A) x^{(k)} + (D+L)^{-1} b$$

De la misma manera, introducimos el método SOR, successive over-relaxation method.

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right] + (1-\omega) x_i^{(k)}, \quad i = 1, \ldots, n$$

Para $\omega = 1$, coincide con el M. de Gauss-Seidel.

En particular, si $\omega \in (0,1)$ se llama sub-relajado si $\omega > 1$ se llama sobre-relajado.

SOR.

A, b, x0, nmax, tol, ω

```
iter = 0
r = b - A x0
r0 = norm(r)
err = norm(r)
x old = x0,
while err > tol & iter < nmax
    iter = iter + 1
        for i = 1 : n
         s = 0
```

```
for j = 1 : i-1
S = S + A(i,j) * x(j)
end
for j = i+1 : n
S = S + A(i,j) * xold(j)
end
x(i) = ω (b(i) - S) / A(i,i) + (1-ω) * xold(i),
end
xold = x,
r = b - Ax,
err = norm(r) / r0
end.
```

Teo Si el método de Jacobi es convergente, entonces el método JOR converge si $0 < \omega < 1$.

Demost. Tenemos en cuenta que el método JOR se escribe como:

$$B_{J_\omega} = \omega B_J + (1-\omega) I.$$

donde $B_J = I - D^{-1}A$.

Sean x_k y λ_k los vectores y valores propios de B_J, entonces.

$$B_{J_w} X_k = (\omega B_J + (1-\omega) I) X_k$$

$$= \omega B_J X_k + (1-\omega) X_k$$

$$= \omega \lambda_k X_k + (1-\omega) X_k$$

$$= \underbrace{(\omega \lambda_k + (1-\omega))}_{\mu_k} X_k$$

Así, los val. p. de B_{J_w} son $\mu_k = \omega\lambda_k + (1-\omega)$

$k = 1, \dots, n$

Usando la fórmula de Euler para la representación de números complejos, tenemos que, $\lambda_k = r_k e^{i\theta_k}$, $\rho(B_J) < 1$

$\max_k |\lambda_k| < 1$

Así:

$$|\mu_k|^2 = |\omega \lambda_k + (1-\omega)|^2$$

$$= |\omega r_k e^{i\theta_k} + (1-\omega)|^2$$

$$= |\omega^2 r_k^2 (e^{i\theta_k})^2| + 2\omega r_k e^{i\theta_k}(1-\omega) + (1-\omega)^2$$

$$= \omega^2 r_k^2 + 2\omega r_k (1-\omega) + (1-\omega)^2 \quad \text{(tomamos la parte real)}$$

$$= (\omega r_k + 1 - \omega)^2$$

$$= ((r_k - 1)\omega + 1)^2 < 1$$

$\Rightarrow$ $\omega > 0$ si $r_k < 1$.

$\dots < 1$

$\Rightarrow$ $\omega > 0$ si $\forall k < 1$.
$\omega \leq 1$

Teo: Para cualquier $\omega \in \mathbb{R}$ tenemos que $\rho(B(\omega)) \geq |\omega - 1|$, por lo tanto, el método SOR falla en converger si $\omega \leq 0$ ó $\omega \geq 2$. (ver cuaderno).

Métodos de descenso

martes, 9 de febrero de 2021 12:02

$$Ax=b$$

↳ Métodos directo (A)

↳ Fact. A

↳ Métodos it. clásicos. $\phi(x^*)=x^*$, $(x_0, x_1, \dots)$

→ $x^{k+1} = Bx^k + C$ converge $\Leftrightarrow \rho(B) < 1$

↳ R.

↳ J

↳ G-S.

↳ Métodos de descenso (it).

$(x_0, x_1, \dots, \dots)$

Sean $A \in \mathbb{R}^{n\times n}$ una matriz s.d.p y $b \in \mathbb{R}^n$, el problema $Ax-b=0$ se puede formular como un problema de optimización:

$$\min_{x\in\mathbb{R}^n} \phi(x) = \frac{1}{2} x^T A x - b^T x.$$

verifiquemos ésta equivalencia. veamos que si x es el mínimo para ϕ (ϕ es una función convexa estricta, cuadrática y diferenciable). entonces x satisface la ec. lineal $Ax=b$.

Usemos la def. de la derivada direccional, ya que nuestra función ϕ es diferenciable.

$$(\phi'(x), y) = \phi'(x)y = \lim_{t\to 0^+} \frac{\phi(x+ty)-\phi(x)}{t} \quad \forall y \in \mathbb{R}^n$$

$$\phi(x+ty)-\phi(x) = \frac{1}{2}(x+ty)^T A(x+ty) - b^T(x+ty)$$

$$- \frac{1}{2} x^T A x + b^T x$$

$$= \frac{1}{2}(x+ty)^T A(x+ty) - b^T(x+ty)$$

$$= \frac{1}{2}x^TAx + tx^TAy + \frac{1}{2}t^2y^TAy - b^Tx - tb^Ty$$

$$= tx^TAy + \frac{1}{2}t^2y^TAy - tb^Ty$$

$$= \left(tx^TA + \frac{1}{2}t^2y^TA - tb^T\right)y \qquad \forall y \in \mathbb{R}^n$$

$$\lim_{t\to 0^+} \frac{\phi(x+ty)-\phi(x)}{t} = \lim_{t\to 0^+} \frac{\left(tx^TA - \frac{1}{2}t^2y^TA - tb^T\right)y}{t}$$

$$= \lim_{t\to 0^+} \left(x^TA - \frac{1}{2}ty^TA - b^T\right)y$$

$$= (x^TA - b^T)y \qquad \forall y \in \mathbb{R}^n$$

$$\phi'(x)y = (\phi'(x), y) = (x^TA - b^T)y = \nabla\phi(x)^T y$$

$$\phi'(x) = x^TA - b^T$$

$$\nabla\phi(x)^T = x^TA - b^T$$

Tenemos entonces que el mínimo satisface

$$\phi'(x)y = 0 \quad \forall y \in \mathbb{R}^n,$$

es decir $x^TA - b^T = 0$

$\Rightarrow Ax - b = 0$, pues A es simétrica.

$\Rightarrow$ Si x es mín ϕ $\Rightarrow$ x satisface $Ax - b = 0$

Recíprocamente, si x satisface $Ax = b$, entonces x es mínimo de ϕ.

$\forall y \in \mathbb{R}^n$

es mínimo de ϕ.

$$Ax=b \Rightarrow \phi(x)-\phi(y)= \qquad \forall y\in\mathbb{R}^n$$

$$=\frac{1}{2}x^TAx - b^Tx - \frac{1}{2}y^TAy + b^Ty$$

$$=+\frac{1}{2}x^TAx - x^TAx - \frac{1}{2}y^TAy + x^TAy$$

$$=-\frac{1}{2}x^TAx - \frac{1}{2}y^TAy + x^TAy$$

$$= -\frac{1}{2}(x-y)^TA(x-y) < 0 \quad \forall y\in\mathbb{R}^n$$

$$= -\frac{1}{2}z^TAz < 0$$

$$\Rightarrow \phi(x)-\phi(y) < 0 \quad \forall y\in\mathbb{R}^n$$

$$\phi(x) < \phi(y) \quad \forall y\in\mathbb{R}^n,\ y\neq x$$

$\Rightarrow x$ es mínimo.

Es decir, x es un mínimo global de ϕ. En consecuencia queda probada la equivalencia del sistema $Ax=b$ y el problema de opt.

$$Ax=b \iff \min_{x\in\mathbb{R}^n} \frac{1}{2}x^TAx - b^Tx$$

Consideremos ahora el sig. método iterativo expresado de la sig manera:

$$x^{k+1} = x^k + \alpha_k p^k,$$

donde α_k es el paso tomado en cada iteración

p_k es la dirección.

En este sentidos, x_k y p_k se escogen de manera que en cada paso nos acerquemos al mínimo de ϕ, ie., $\phi(x^{k+1}) < \phi(x^k)$.

Método del descenso más profundo.

El método del descenso más profundo consiste en tomar como dirección precisamente aquella en la cual la función ϕ decrece más rápido, es decir,

$$p^{(k)} = -\nabla\phi(x^k)$$
$$= -(Ax - b)$$
$$= b - Ax$$
$$= r^{(k)}$$

Hay que recordar que el residuo es una medida de cuán lejos estamos de satisfacer la ec $Ax = b$. Para poder determinar el valor del paso α_k debemos pensar que este paso se debe tomar en la dirección p^k de tal manera que minimice ϕ. Por lo tanto construimos una función que dependa del paso.

$$\min_{\alpha} f(\alpha) = \phi(x^k + \alpha p^k) = \phi(x^{k+1})$$

El óptimo α que queremos encontrar lo denotaremos por α_k y debe satisfacer:

$$0 = \frac{df(\alpha_k)}{d\alpha} = f'(\alpha_k) = \phi'(x^k + \alpha_k p^k)\, p^k$$

$$0 = \frac{df(\alpha_k)}{d\alpha} = f'(\alpha_k) = \phi'(x^k + \alpha_k p^k)\, p^k$$

$$= \phi'(x^{k+1})\, r^k$$

$$= r^{k+1\,T} r^k$$

$$\Rightarrow \boxed{0 = r^{k+1\,T} r^k}$$

Los residuos son ortogonales.

$$0 = r^{k+1\,T} r^k = (b - Ax^{k+1})^T r^k$$

$$= (b - A(x^k + \alpha_k p^k))^T r^k$$

$$= (b - Ax^k - A\alpha_k p^k)^T r^k$$

$$= (b - Ax^k)^T r^k - \alpha_k p^{k\,T} A r^k$$

$$= r^{k\,T} r^k - \alpha_k r^{(k)\,T} A r^k$$

$$\Rightarrow \alpha_k = \frac{r^{k\,T} r^k}{r_k^T A r^k}$$

Por tanto podemos escribir el algoritmo para el método del descenso más profundo.

Método del gradiente

$[x] = \text{Gradiente}(A_{sdp}, b, x^0)$

$k = 0$

$r^{(0)} = b - Ax^{(0)}$

$err = 1$

while err > tol (1e-4, 1e-5, 1e-6)

$r^{(k)} = b - Ax^{(k)}$

$\alpha = r^{k\,T} r^k$

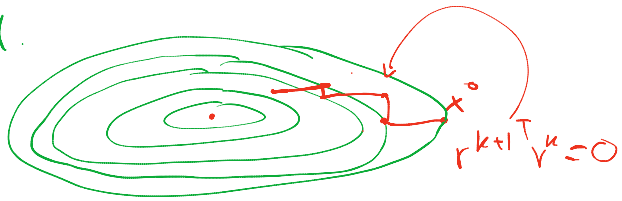
$r^{(k)} = b - Ax^{(k)}$

$\alpha_k = \dfrac{r^{kT} r^k}{r^{kT} A r^k}$

$x^{(k+1)} = x^{(k)} + \alpha_k p^k$

$err = norm(r^k)$

end.



Método del Gradiente Conjugado

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x - b^T x$$

El algoritmo del gradiente conjugado (GC - CG) tiene ciertas semejanzas con el algoritmo del descenso más profundo. El GC sirve para resolver sistemas lineales $Ax = b$, donde $A \in \mathbb{R}^{n \times n}$ es una matriz simétrica definida positiva y $b \in \mathbb{R}^n$. Este algoritmo es más eficiente que el método del descenso más profundo. Una de las características del GC es generar un conjunto de direcciones conjugadas.

Definición: Sean p y q dos vectores de $\mathbb{R}^n$. Se dice que p y q son A-ortogonales o conjugados si

$$p^T A q = 0$$

$$p^T A q = 0$$

Es decir, p y q son A-ortogonales si son ortogonales con respecto al producto escalar inducido por la matriz A.

$$\left(\langle p, q \rangle_A = p^T A q \right)$$

Una de las características del método del GC es que genera un conjunto de vectores $\{ p^{(0)}, p^{(1)}, \dots, p^{(n-1)} \}$ con la propiedad de ser A-ortogonales, ie,:

$$\boxed{p^{(i)T} A p^{(j)} = 0} \quad \forall i \neq j .$$

Esta propiedad hace que el conjunto $\{ p^{(0)}, p^{(1)}, \dots, p^{(n-1)} \}$ sea linealmente independiente.

En efect, para escalares $\alpha_1, \dots, \alpha_{n-1}$

$$0 = \sum_{i=0}^{n-1} \alpha_i p^i \Rightarrow 0 = A \left(\sum_{i=0}^{n-1} \alpha_i p^{(i)} \right)$$

$$\Rightarrow 0 = p^{(j)T} A \left(\sum_{i=0}^{n-1} \alpha_i p^{(i)} \right)$$

$$\Rightarrow 0 = \sum_{i=0}^{n-1} \alpha_i p^{(j)T} A p^{(i)}$$

de donde, por la A-ortogonalidad de los $p^{(i)}$, y como A es def. positiva, la anterior identidad es posible si $\alpha_i = 0$ para $i = 0, \dots, n-1$.

anterior identidad es posible $\alpha_i = 0$

para $i = 0, \dots, n-1$.

$$Ax = b \iff \min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x - b^T x$$

Estas direcciones serán utilizadas como direcciones de descenso para minimizar ϕ usando el esquema iterativo de descenso

$$x^{k+1} = x^k + \alpha_k p^k$$

donde la longitud de cada paso está dada por

$$\rightarrow \alpha_k = \frac{r^{k^T} p^k}{p^{(k)^T} A p^{(k)}} \; ?$$

Vamos a ver que el esquema $x^{k+1} = x^k + \alpha_k p^k$ va a converger a la solución $\bar{x}$ del sistema lineal $Ax = b$.

Por independencia lineal tenemos que

$$\text{span}(p^{(0)}, p^{(1)}, \dots, p^{(n-1)}) = \mathbb{R}^n$$

$\Rightarrow$ Existen $(\sigma_0, \sigma_1, \dots, \sigma_{n-1})$:

$$\boxed{\bar{x} - x_0 = \sum_{i=0}^{n-1} \sigma_i p^{(i)}}$$

$i=0$

$$x^{k+1} = x^k + \alpha_k p_k$$

$$x^{k+1} - x^k = \alpha_k p_k$$

$$x^{k+1} - (x^{k-1} + \alpha_{k-1} p_{k-1}) = \alpha_k p_k$$

$$x^{k+1} - ((x^{k-2} + \alpha_{k-2} p_{k-2}) + \alpha_{k-1} p_{k-1}) = \alpha_k p_k$$

$$\vdots$$

$$\boxed{x^{k+1} - x^0} = \sum_{i=0}^{k} \alpha_k p_k$$

σ_i

Tenemos

$$\bar{x} - x_0 = \sum_{i=1}^{n-1} \sigma_i p^{(i)} \quad (*)$$

Multiplicamos por $p^{(k)^T} A$ para despejar σ_k

$$p^{k^T} A (\bar{x} - x_0) = \sum_{i=0}^{n-1} \sigma_i p^{(k)^T} A p^{(i)}$$

$$= \sigma_k p^{k^T} A p^k$$

$$\sigma_k = \frac{p^{k^T} A (\bar{x} - x_0)}{p^{k^T} A p^k}$$

σ_k depende de $\bar{x}$ que es desconocida, sin embargo aprovecharemos la A-ortogonalidad para encontrar una expresión más conveniente.

$$x^k = x^{k-1} + \alpha_{k-1} p^{k-1}$$

$$= x^{k-2} + \alpha_{k-2} p^{k-2} + \alpha_{k-1} p^{k-1}$$

$$\vdots$$

$$= x^{(0)} + \alpha_0 p^0 + \alpha_1 p^1 + \cdots + \alpha_{k-1} p^{k-1}$$

Multiplicamos por $p^{k^T} A$ y tenemos:

$$p^{k^T} A x^k = p^{k^T} A x^{(0)} + \alpha_0 p^{k^T} A p^{(0)} + \cdots + \alpha_{k-1} p^{k^T} A p^{k-1}$$

$$\Rightarrow \boxed{p^{k^T} A x^k = p^{k^T} A x^{(0)}}$$

Entonces, de esta identidad tenemos:

$$p^{k^T} A(\bar{x} - x^0) = p^{k^T} A(\bar{x} - x^k) + p^{k^T} A(x^k - x^0)$$

$$= p^{k^T} A \bar{x} - p^{k^T} A x^k$$

$$= p^{k^T} b - p^{k^T} A x^k$$

$$= p^{k^T} (b - A x^k)$$

$$= p^{k^T} r^k$$

Así $\alpha_k = \dfrac{r^{k^T} p^k}{p^{k^T} A p^k} = \alpha_k$ ✓

Lo cual indica que los pasos y las direcciones son aquellas de la combinación lineal de (*). Es decir el método converge en n iteraciones. Sin embargo,

lineal de (*). Es decir el método converge en n iteraciones. Sin embargo, en la práctica esto no sucede y el GC se comporta como un método iterativo por los efectos de redondeo y truncamiento, lo que impide la convergencia teórica en n pasos.

2 Métodos de descenso

$\min \frac{1}{2} x^T A x - b^T x$

- Gradiente (Descenso prof)
  - $x^{k+1} = x^k + \alpha_k p^k$
- Grad. Conjugado
  - $\{p^0, p^{(1)}, \ldots, p^{(n-1)}\}$
    - span ( ) $= \mathbb{R}^n$
    - A-conjugados.

$\bar{x} - x_0 = \sum_{i=0}^{n-1} \alpha_i p_i$ ✓ ?

Prueba del 22 de febrero

- Mínimos cuadrados
- Met. It. clásicos
- Método del Gradiente

Método GC.

$$x^{k+1} = x^k + \alpha^k p^k$$

$$\bar{x} - x_0 = \sum_{i=0}^{n-1} \alpha_k p^k$$

$$\text{span}(p^0, p^1, \ldots, p^{n-1}) = \mathbb{R}^n$$

generar

Nos falta construir las direcciones A-ortogonales. Vamos a aprovechar las propiedades de los residuos para generar $\{p^{(0)}, p^{(1)}, \dots, p^{(n-1)}\}$.

Definimos la siguiente función:

$$h : \mathbb{R}^k \to \mathbb{R}$$

$$h(\underbrace{\sigma}_{(\sigma_0, \sigma_1, \dots, \sigma_{k-1})}) = \phi(x^0 + \sigma_0 p^0 + \sigma_1 p^1 + \dots + \sigma_{k-1} p^{k-1})$$

Esta función es convexa y continua, por lo tanto, tiene un único mínimo que satisface:

$$0 = \nabla h(\bar{\sigma})$$

equivale a que:

$$0 = \frac{\partial h(\bar{\sigma})}{\partial \sigma_i} = \nabla \phi(\underbrace{x^0 + \bar{\sigma}_0 p^{(0)} + \dots + \bar{\sigma}_{k-1} p^{k-1}}_{\tilde{x}})^T p^{(i)}$$

$$= (b - A\tilde{x})^T p^{(i)}$$

$$= r(\tilde{x})^T p^i$$

Concluimos que $\tilde{x} = x^{(0)} + \sum_{i=0}^{k-1} \bar{\sigma}_i p^{(i)}$ minimiza ϕ en $x^{(0)} + \text{span}(p^{(0)}, p^{(1)}, \dots, p^{k-1})$ ssi

$r(\tilde{x})^T p^{(i)} = 0, \quad i = 0, \dots, k-1.$

Lema: En el k-ésimo paso del método de GC se satisface la siguiente relación de ortogonalidad:

$$(*)\quad r^{k\,T} p^{(i)} = 0 \quad \text{para } i = 0, \dots, k-1$$

Demost. Para $x^{(1)} = x^{(0)} + \alpha_0 p^{(0)}$ se tiene

$h(\alpha_0) = \phi(x^{(0)} + \alpha_0 p^{(0)})$, por lo tanto, si α_0 es el mínimo de h, entonces

$$\begin{aligned} h'(\alpha_0) = 0 &= \phi'(x^{(1)})^T p^{(0)} \\ &= (b - Ax^{(1)})^T p^{(0)} \\ &= r^{(1)\,T} p^{(0)} \end{aligned}$$

Así, $(*)$ se cumple para $k = 1$.

Supongamos que se cumple para $i = 1, \dots, k-2$, i.e., $r^{k-1\,T} p^{(i)} = 0$, para $i = 1, \dots, k-2$.

Como $x^k = x^{k-1} + \alpha_{k-1} p^{k-1}$, el residuo es

$$\begin{aligned} r^k &= b - Ax^k \\ &= b - A(x^{k-1} + \alpha_{k-1} p^{k-1}) \\ &= b - Ax^{k-1} - \alpha_{k-1} A p^{k-1} \\ &= r^{k-1} - \alpha_{k-1} A p^{k-1}. \end{aligned}$$

Multiplicamos la identidad previa por $p^{(i)\,T}$ y deducimos:

$$p^{(i)T} r^{(k)} = \underbrace{p^{(i)T} r^{(k-1)}}_{=0} - \alpha_{k-1} \underbrace{p^{(i)T} A p^{(k-1)}}_{=0}$$

$$= 0 \quad \text{para } i = 0, \dots, k-2.$$

Falta probar para $i = k-1$. Así:

$$p^{(k-1)T} r^{k} = p^{k-1T}(b - Ax^{k})$$

$$= p^{k-1T}(b - Ax^{k-1} - \alpha_{k-1} A p^{k-1})$$

$$= p^{k-1T} r^{k-1} - \alpha_{k-1} p^{k-1T} A p^{k-1}$$

$$= p^{k-1T} r^{k-1} - \frac{r^{k-1T} p^{k-1}}{p^{k-1T} A p^{k-1}} p^{k-1T} A p^{k-1}$$

$$= p^{k-1T} r^{k-1} - r^{k-1T} p^{k-1}$$

$$= 0$$

El Lema anterior es central en el método del GC ya que nos permite construir el conjunto de direcciones (A-ortogonales) de manera eficiente.

Buscaremos un coeficiente β_k que sirva para calcular la nueva dirección a partir de la anterior.

$$p^{(k)} = r^{(k)} + \beta_k p^{(k-1)}$$

$$p^{(k)} = r^{(k)} + \beta_k p^{(k-1)}$$

Los coeficientes β_k a determinar deben cumplir con el requerimiento que los $p^{(k)}$ sean A-ortogonales.

$$p^{k-1^T} A p^k = p^{k-1^T} A r^k + \beta_k p^{k-1^T} A p^{k-1} = 0$$

Despejamos β_k y tenemos:

$$\beta_k = - \frac{p^{k-1^T} A r^k}{p^{k-1^T} A p^k}$$

Algoritmo preliminar del GC.

$[x] = GC(A_{(s.d.p)}, b, x^{(0)})$

$k = 0$

$p^{(0)} = r^{(0)} = b - Ax^{(0)}$

while $\|r^{(k)}\| > tol$

$$\alpha_k = \frac{r^{k^T} p^k}{p^{k^T} A p^k}$$

$$x^{k+1} = x^k + \alpha_k p^k$$

$$r^{k+1} = b - Ax^{k+1}$$

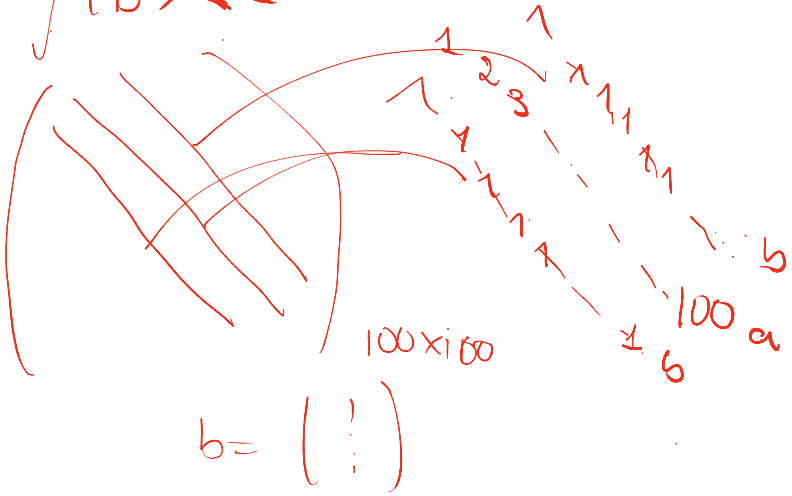
$$\rightarrow \beta_{k+1} = - \frac{p^{k^T} A r^{k+1}}{p^{k^T} A p^k}$$

$$\rightarrow p^{k+1} = r^{k+1} + \beta_{k+1} p^k$$

→ $p^{k+1} = r^{k+1} + \beta_{k+1} p^k$

end.

$\rho(B) < 1 \iff$ MIE_converge



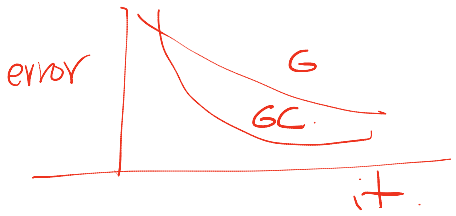
$b = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

→ spdiags [b a b], -1, 0, 1, n, n]

diag

→ 100 iteraciones de los métodos

→ error $\frac{\|b - Ax\|}{\|r_0\|}$



$r^k \in K(r^0, k) = \text{Span}\{r^{(0)}, Ar^{(0)}, \dots, A^k r^{(0)}\}$

Ahora estudiaremos otras propiedades del Alg. del GC.
Veremos que los residuos $r^{(i)}$ son mutuamente ortogonales, segundo, cada dirección de búsqueda $p^{(i)}$ y los residuos $r^{(i)}$ se encuentran contenidos en el subespacio de Krylov de orden k para $r^{(0)}$.

Los subespacios de Krylov tienen la característica de que son generados recursivamente.

$$\mathcal{K}(r^{(0)}, k) = \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^k r^{(0)}\}$$

Teo . Supongamos que la k-ésima iteración del algoritmo del GC no coincide con la solución $\bar{x}$, entonces las sig. propiedades se satisfacen: $r^{(0)}$. n pasos

$$\begin{cases} r^{k^T} r^{(i)} = 0 \quad \forall i = 0, \dots, k-1 \\ \text{span}\{r^{(0)}, r^{(1)}, \dots, r^{(k)}\} = \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^k r^{(0)}\} \\ \text{span}\{p^{(0)}, p^{(1)}, \dots, p^{(k)}\} = \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^k r^{(0)}\} \\ p^{(k)^T} A p^{(i)} = 0 \quad \forall i = 0, \dots, k-1 \end{cases}$$

Demost. Pendiente.

Gracias a este Teorema modificamos el Alg. del G.C. y lo escribimos de la siguiente manera:

$$\alpha_k = \frac{r^{(k)^T} p^{(k)}}{p^{(k)^T} A p^{(k)}} = \frac{r^{(k)^T} r^{(k)}}{p^{(k)^T} A p^{(k)}} + \beta_k \frac{r^{(k)^T} p^{k-1} \; 0}{p^{k^T} A p^k}$$

$$\alpha_k = \frac{r^{k^T} r^k}{p^{k^T} A p^k}$$

Ahora $r^{k+1} = r^k - \alpha_k A p^{(k)}$

$$r^{k+1T} r^{k+1} = \underbrace{r^{k+1T} r^k}_{0} - \frac{r^{kT} r^k}{p^{kT} A p^k} r^{k+1T} A p^k$$

$$= -r^{kT} r^k \beta_{k+1}$$

$$\Rightarrow \boxed{\beta_{k+1} = -\frac{r^{k+1T} r^{k+1}}{r^{kT} r^k}} \leftarrow$$

Con estas modificaciones, reescribimos el algoritmo:

$$p^{(0)} = r^{(0)}$$

$$\alpha_k = \frac{r^{kT} r^k}{p^{kT} A p^k}$$

$$x^{k+1} = x^k + \alpha_k p^k$$

$$r^{k+1} = r^k - \alpha_k A p^k$$

$$\beta_{k+1} = \frac{r^{k+1T} r^{k+1}}{r^{kT} r^k}$$

$$p^{k+1} = r^{k+1} - \beta_{k+1} p^k$$

Precondicionamiento

El problema $Ax=b$ se puede reformular de la siguiente manera:

$$\boxed{\quad} \quad (1)$$

la sig. manera:

$$P^{-1}Ax = P^{-1}b \qquad (1)$$

donde el problema (1) es un sistema precondicionado donde P es la matriz precondicionadora por la izquierda. Existen precondicionadores por la derecha y centrados:

$$AP^{-1}y = b, \quad y = Px$$

y

$$P_L^{-1} A P_R^{-1} y = P_L^{-1} b, \quad y = P_R x$$

Los métodos it. clásicos corresponden a iteraciones de punto fijo que pueden ser precondicionados por la izquierda. Por ejemplo, tenemos que un método it. tiene el sig. esquema:

$$x^{k+1} = Bx_k + c$$

$$B_J = I - D^{-1}A$$

$$B_{GS} = I - (D+L)^{-1}A$$

Para los métodos de Jacobi y G-S tenemos:

$$A = M - N \longleftarrow$$

$$(Ax = b)$$

$$(M-N)x = b$$

$$Mx = b - Nx$$

$$Mx = b - Nx$$

$$x = \underbrace{M^{-1}N}_{B}x + \underbrace{M^{-1}b}_{c}$$

Por lo tanto, la iteración $x_{k+1} = Bx_k + c$ se puede ver como una técnica para resolver el sistema:

$$(I - B)x = c$$

$$(I - B)x = M^{-1}b \quad | \; b = Ax$$

$$(I - B)x = M^{-1}Ax$$

Así $B = I - M^{-1}A$ y podemos escribir el sistema de la siguiente manera:

$$\boxed{M^{-1}Ax = M^{-1}b}$$

Por lo tanto, vemos que la matriz M sirve como matriz precondicionadora para los métodos iterativos clásicos:

$$M_J = D$$

$$M_{GS} = D + L$$

$$M_{JOR} = \frac{1}{\omega}(D + \omega L)$$

$$M_{SOR} = \frac{1}{\omega(2-\omega)}(D + \omega L)D^{-1}(D + \omega U)$$

Note que el sistema precondicionado puede ser un sist. lleno o denso, es decir, no existe razón para que la matriz M^{-1} sea sparse, incluso si M es sparse. Esto limita el número de técnicas que pueden aplicarse al sist. precondicionado.

La mayoría de los métodos it. usados sólo requieren productos de matriz por vector.

En este caso, para calcular

$$w = M^{-1}Av$$ para cualquier vector v

primero requerimos calcular $r = Av$ y luego resolver

$$w = M^{-1}r, \text{ i.e., } Mw = r$$

$$\Rightarrow \boxed{\begin{aligned} r &= Av \\ w &= M^{-1}r \end{aligned}}$$

En algunos casos, resulta ventajoso explotar que $A = M - N$ y calcular

$$w = M^{-1}Av$$

$$w = (I - M^{-1}N)v$$

$$\Rightarrow w = v - M^{-1}Nv$$

Así tenemos el procedimiento

Así tenemos el procedimiento

$$r = Nv$$
$$w_1 = M^{-1} r$$
$$w = v - w_1$$

La matriz N es más sparse que A y el producto matriz por vector Nv es menos caro que Av.

Revisión de precondicionadores:

⊙ Precondicionadores diagonales: Escoger P (en el caso general) como la diagonal de A es en general efectivo si A es s.d.p. En el caso no-simétrico, se puede tomar:

$$P_{ii} = \left(\sum_{j=1}^{n} a_{ij}^2 \right)^{1/2}$$

⊙ Factorización LU incompleta y Fact. Cholesky incompleta.

La fact. incompleta de A es un proceso que calcula $P = L_{in} U_{in}$, donde L_{in} es Δ y U_{in} es ∇. Estas matrices son aproximaciones de las matrices L y U de la f. LU exacta. Se escogen de tal manera que

de las matrices L y U de ...

Se escogen de tal manera que

$R = A - L_{in} U_{in}$ satisfaga ciertos requerimientos como tener ceros en entradas específicas de la matriz.

En el caso de matrices s.d.p, uno puede obtener un mejor precondicionador con la Fact. de Cholesky modificada incompleta.

Trabajo: G-C Precondicionado.

$$\begin{cases} -\Delta u = f & \text{en } \Omega \\ u = 0 & \text{en } \partial\Omega. \end{cases}$$

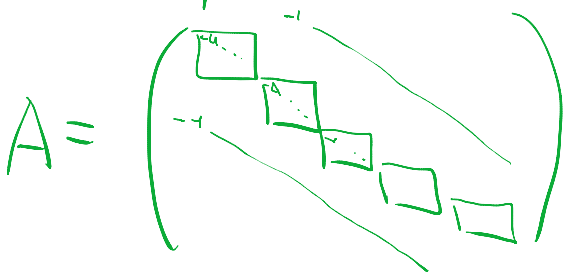
$$\underline{Au = b}$$

→ Usar el Método del gradient precondicionado.

→ Matriz A

G = numgrid ('S', n)

A = delsq (G)



$$A = \begin{pmatrix} \boxed{\begin{matrix} -4 & \\ & \ddots \end{matrix}} & & & & \\ & \boxed{\begin{matrix} -4 & \\ & \ddots \end{matrix}} & & & \\ & & \boxed{\begin{matrix} -4 & \\ & \ddots \end{matrix}} & & \\ & & & \boxed{} & \\ & & & & \boxed{} \end{pmatrix}$$

$b = (10)$

$tol = 10^{-10}$

$x^{(0)} = 0^T$

$nmaxit = 200$

→ $n = 12, 22, 27$ $(100, 400, 625)$

Precondicionador

$P = L^T L$ (factorización incompleta de Cholesky)

consultar

Gradiente conjugado Precondicionado

$P = M = LL^T$ ←

→ $\left\{ \begin{array}{l} L^{-1} A L^{-T} u = L^{-1} b \\ x = L^{-T} u \end{array} \right\}$

Algoritmo del Grad. conj. Precond.

$r_0 = b - Ax_0$

$Pz_0 = r_0$ (resolvemos el sist. lineal con Cholesky incompleto)

$p_0 = z_0$

while criterio de parada no se satisfaga $(\|r\| > tol)$

$$\alpha_k = \frac{r_k^T z_k}{p_k^T A p_k}$$

$$x^{k+1} = x^k + \alpha_k p_k$$

$$r^{k+1} = r^k - \alpha^k A p^k$$

→ $P z^{k+1} = r^{k+1}$ (resolvemos con Cholesky incompleto)

$$\beta_k = \frac{r^{k+1^T} z^{k+1}}{r^{k^T} z^k} \leftarrow$$

$$p_{k+1} = z^{k+1} + \beta_k p^k$$

cual

Expo:

→ Presentar el problema

$$-\Delta u = f \text{ en } \Omega$$
$$u = 0 \text{ en } \partial\Omega$$

problema de Poisson.

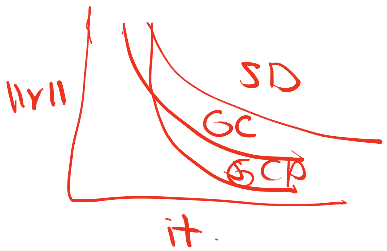
→ Presentar el Algoritmo G.C.P

→ Presentar el Alg. PC-Incompleta $Ax = b$ usdp

→ Presentar los resultados

it / $\|r\|$ / tiempo

n	SD	GC	GCP
12			
22			
27 !			



→ Bibliografía $\frac{\|r\|}{\|r_0\|} = \frac{\|b - Ax\|}{\|b - Ax_0\|}$

→ Cuaternari

→ Iterative Methods for Sparse Linear Systems

Yousef Saad

→ Lab /10 (30%) — Lab 1 /1, Lab 2 /3, Expo /6

→ Teo /10 (70%) → /5, /5

Teo Supongamos que la k-ésima it. del Alg. 6.c no coincide con la solución $\bar{x}$, entonces las sig. propiedades se satisfacen:

$$r^{(k)T} r^{(i)} = 0 \quad \forall i = 0, \dots, k-1 \qquad (1)$$

$$\text{span}\{r^{(0)}, \dots, r^{(k)}\} = \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^k r^{(0)}\} \qquad (2)$$

$$\text{span}\{p^{(0)}, \dots, p^{(k)}\} = \text{span}\{r^{(0)}, \dots, A^k r^{(0)}\} \qquad (3)$$

$$p^{(k)T} A p^{(i)} = 0, \quad \forall i = 0, \dots, k-1 \qquad (4)$$

Record: $r^{k^T} p^{(i)} = 0$, para $i = 0, \dots, k-1$ $(\star)$

Demost: Procedemos por inducción. Para (2) y (3) se satisface trivialmente para $k=0$. Para (1), se cumple para $k=1$ gracias a $(\star)$. Y (4) se tiene que para $k=1$:

$$p^{(1)T} A p^{(0)} = \left(r^1 + \beta_1 p^{(0)}\right)^T A p^{(0)}$$

$$= r^{1T} A p^{(0)} - \frac{p^{(0)T} A r^{(1)}}{p^{(0)T} A p^{(0)}} p^{(0)T} A p^{(0)}$$

$$= 0$$

Ahora, tenemos que $r^{k+1} = r^k - \alpha_k A p^k$, i.e, r^{k+1} es combinación lineal de r^k y Ap^k. Sin embargo, consideremos la hipótesis inductiva dada por (2)-(4) y observemos que:

es combinación lineal de r^m y Ap^k. Sin embargo, consideremos la hipótesis inductiva dada por (2)-(4). Y observemos que:

$$r^k \in \text{span}\{r^{(0)}, \ldots, A^k r^{(0)}\}$$

$$p^{(k)} \in \text{span}\{r^{(0)}, \ldots, A^k r^{(0)}\},$$

de donde se tiene que Ap^k satisface

$$Ap^k \in \text{span}\{Ar^{(0)}, \ldots, A^{k+1} r^{(0)}\} \subseteq \mathcal{K}(r^0, k+1).$$

En consecuencia

$$r^{k+1} \in \text{span}\{r^{(0)}, \ldots, A^{k+1} r^{(0)}\}, \text{ y}$$

combinando con nuestra h. inductiva (2), tenemos

$$\text{span}\{r^{(0)}, r^{(1)}, \ldots, r^k, r^{k+1}\} \subset \text{span}\{r^{(0)}, \ldots, A^{k+1} r^{(0)}\}.$$

Para concluir la inclusión en el otro sentido (⊃), utilizamos la h. inductiva (3).

$$A^{k+1} r^{(0)} = A A^k r^{(0)} \in \text{span}\{Ar^{(0)}, \ldots, A^k r^{(0)}, A^{k+1} r^{(0)}\}$$

$$= \text{span}\{Ap^{(0)}, \ldots, Ap^{(k)}\}.$$

Por otro lado, $r^{(i+1)} = r^i - \alpha_i Ap^i$ se tiene que

$$Ap^{(i)} = \frac{r^{(i+1)} - r^{(i)}}{\alpha_i}, \text{ para } i = 1, \ldots, k, \text{ de aquí,}$$

$$A^{k+1} r^{(0)} \in \text{span}\{r^{(0)}, \ldots, r^{(k+1)}\}.$$

$A^{k+1} r^{(0)} \in \text{span}\{r^{(0)}, \dots, r^{(k+1)}\}$.

Combinando la última expresión con la H. inductiva se sigue que:

$$\text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{k+1} r^{(0)}\} = \text{span}\{r^{(0)}, \dots, r^{k+1}\}.$$

Así, hemos probado (2). Ahora probaremos (3) para $k+1$.

$$\text{span}\{p^{(0)}, \dots, p^{k}, p^{(k+1)}\} = \text{span}\{p^{(0)}, p^{(1)}, \dots, p^{(k)}, r^{k+1}\}, \quad p^{k+1} = r^{k+1} + \beta_{k+1} p^{(k)}$$

$$= \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{k} r^{(0)}, r^{k+1}\}$$

$$= \text{span}\{r^{(0)}, r^{(1)}, \dots, r^{(k)}, r^{k+1}\}$$

$$= \text{span}\{r^{(0)}, Ar^{(0)}, \dots, A^{k+1} r^{(0)}\} \quad \text{(por (2) en } k+1)$$

Así, queda probado (3). Falta por probar que las direcciones $p^{(i)}$ son A- ortogonales para $k+1$.

Por inducción, $p^{k+1} = r^{k+1} + \beta_{k+1} p^{k}$, multiplicamos por $Ap^{(i)}$:

$$p^{k+1\,T} A p^{(i)} = (r^{k+1} + \beta_{k+1} p_k)^T A p^{(i)}$$

$$= r^{k+1\,T} A p^{(i)} - \frac{p^{k\,T} A r^{k+1}}{p^{k\,T} A p^{k}} p^{k\,T} A p^{(i)}$$

Si $i = k$, tenemos $p^{k+1\,T} A p^{k} = 0$. Por nuestra hipótesis de inducción, las direcciones $p^{(0)}, \dots, p^{(k)}$

son A-ortogonales, así, de (*), tenemos:

$$r^{k+1^T} p^{(i)} = r^{k^T} p^{(i)} - \alpha_k \overbrace{p^{k^T} A p^{(i)}}^{\text{H. ind.}} = 0 \quad , \text{para } i = 0, \dots, k-1 \qquad (\square)$$

Por otro lado, para $i = 1, \dots, k-1$, de (3) tenemos:

$$A p^{(i)} \in A \cdot \text{span}\{ r^{(0)}, Ar^{(0)}, \dots, A^i r^{(0)} \}$$

$$= \text{span}\{ A r^{(0)}, \dots, A^{i+1} r^{(0)} \}$$

$$= \text{span}\{ p^{(0)}, \dots, p^{(i+1)} \} \qquad (2\square)$$

De $(\square)$ y $(2\square)$, tenemos

$$r^{k+1^T} A p^{(i)} = 0 \quad , \; i = 1, \dots, k-1.$$

y así deducimos (4).

Falta probar (1). Por (A) tenemos

- $r^{k^T} p^{(i)} = 0$, para $i = 0, \dots, k-1$ y para cualquier $k = 1, \dots, n-1$, por construcción, tenemos

$$p^{(i)} = r^{(i)} + \beta_i p^{(i-1)} \Rightarrow r^{(i)} = p^{(i)} - \beta_i p^{(i-1)}$$

$$\Rightarrow r^{(i)} \in \text{span}\{ p^{(i)}, p^{(i-1)} \}, \quad \forall i = 1, \dots, k-1$$

- $r^{k^T} (r^i + \beta_i p^{i-1}) = r^{k^T} r^{(i)} + \beta_i r^{k^T} p^{i-1} = 0$

T.

$$\Rightarrow r^{k^T} r^{(i)} = 0, \quad \forall i = 0, \dots, k-1.$$

Gracias a este Teo. pudimos reescribir

$$\beta_{k+1} = \frac{r^{k+1^T} r^{k+1}}{r^{k^T} r^k}.$$

→ Detalles métodos de descenso.

→ Valores y vectores propios.

Convergencia del método del Gradiente.

Supongamos x^* satisface $Ax^* = b$ exactamente.

Ahora, analicemos el error en la iteración k-ésima:

$$R_k = \frac{\phi(x_k) - \phi(x^*)}{\phi(x_{k-1}) - \phi(x^*)} < \beta < 1$$

Si R_k está acotado por β tq. $R_k < \beta < 1$ para un β fijo entonces $\phi(x_k) - \phi(x^*) \to 0$ si $k \to \infty$.

Por lo tanto el método del gradiente converge a x^*.

Analicemos R_k. Por conveniencia expandimos $\phi(x_k)$.

$$\phi(x_k) = \phi(x_{k-1} + \alpha_k p_k)$$

$$= \frac{1}{2}(x_{k-1} + \alpha_k p_k)^T A (x_{k-1} + \alpha_k p_k) - b^T (x_{k-1} + \alpha_k p_k)$$

$$= \frac{1}{2}(x_{k-1} + \alpha_k p_k)^T A (x_{k-1} + \alpha_k p_k) - b^T (x_{k-1} + \alpha_k p_k)$$

$$= \phi(x_{k-1}) + \alpha_k p_k^T A x_{k-1} + \frac{1}{2}\alpha_k^2 p_k^T A p_k - \alpha_k b^T p_k$$

$$= \phi(x_{k-1}) + \alpha_k p_k^T (b - p_k) + \frac{1}{2}\alpha_k^2 p_k^T A p_k - \alpha_k b^T p_k$$

$$= \phi(x_{k-1}) - \alpha_k p_k^T p_k + \frac{1}{2}\alpha_k^2 p_k^T A p_k$$

$$= \phi(x_{k-1}) - \frac{p_k^T p_k}{p_k^T A p_k}(p_k^T p_k) + \frac{1}{2}\left(\frac{p_k^T p_k}{p_k^T A p_k}\right)^2 p_k^T A p_k$$

$$= \phi(x_{k-1}) - \frac{(p_k^T p_k)^2}{2 p_k^T A p_k}$$

Podemos usar esta fórmula para encontrar una expresión alternativa al error R_k.

$$R_k = \frac{\phi(x_{k-1}) - \frac{(p_k^T p_k)^2}{2 p_k^T A p_k} - \phi(x^*)}{\phi(x_{k-1}) - \phi(x^*)}$$

$$= 1 - \frac{(p_k^T p_k)^2}{2 p_k^T A p_k (\phi(x_{k-1}) - \phi(x^*))}$$

Para simplificar la diferencia en el denominador, podemos usar $x^* = A^{-1}b$.

$$\phi(x_{k-1}) - \phi(x^*) = \left[\frac{1}{2} x_{k-1}^T A x_{k-1} - b^T x_{k-1}\right] - \left[\frac{1}{2} x^{*T} b - b^T x^*\right]$$

$$= \frac{1}{2} x_{k-1}^T A x_{k-1} - b^T x_{k-1} + \frac{1}{2} b^T A^{-1} b \quad , x^* = A^{-1}b$$

$$= \frac{1}{2}\left((Ax_{k-1})^T A^{-1} A x_{k-1} - b^T A^{-1} A x_{k-1} + b^T A^{-1} b\right)$$

$$= \frac{1}{2}(Ax_{k-1} - b)^T A^{-1}(Ax_{k-1} - b) = \frac{1}{2}(-p_k)^T A^{-1}(-p_k)$$

$$R_k = 1 - \frac{(p_k^T p_k)^2}{2 p_k^T A p_k \left(\frac{1}{2} p_k^T A^{-1} p_k\right)}$$

$$= 1 - \frac{(p_k^T p_k)^2}{p_k^T A p_k \cdot p_k^T A^{-1} p_k}$$

$$= 1 - \frac{(p_k^T p_k)}{p_k^T A p_k \cdot p_k^T A^{-1} p_k}$$

$$= 1 - \frac{p_k^T p_k}{p_k^T A p_k} \cdot \frac{p_k^T p_k}{p_k^T A^{-1} p_k}$$

$$\leq 1 - \left(\min_{\|p\|=1} \frac{1}{p^T A p} \right) \left(\min_{\|p\|=1} \frac{1}{p^T A^{-1} p} \right)^{-1}$$

$$= 1 - \left(\max_{\|p\|=1} p^T A p \right)^{-1} \left(\max_{\|p\|=1} p^T A^{-1} p \right)^{-1}$$

$$= 1 - \frac{\sigma_{min}}{\sigma_{max}},$$

σ_{min}, σ_{max} son el mínimo y máximos valores singulares, respecto de la matriz A.

$$= 1 - \frac{1}{\text{cond } A_{\|\cdot\|_2}}$$

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \qquad A = U V \Sigma^T$$

$$= \sigma_{max}$$

$$\|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{max}}{\sigma_{min}}$$

$$\|A^{-1}\|_2 = \frac{1}{\sigma_{min}},$$

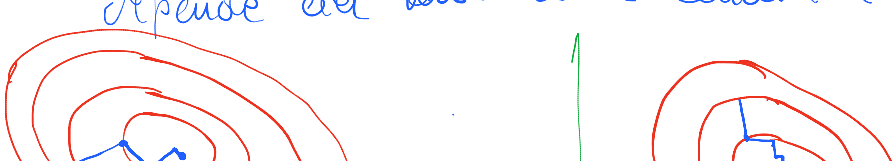
σ_{max} es el valor singular más grande de la matriz A.

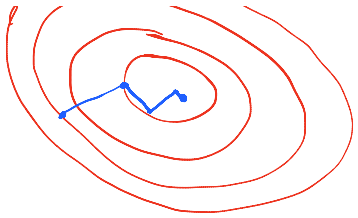
σ_{min} es el " " " pequeño distinto de cero.

$$\boxed{R_k} \leq 1 - \frac{1}{\text{cond } A_{(\|\cdot\|_2)}} \qquad \boxed{\text{cond}(A) \geq 1}$$

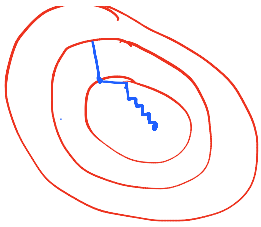
$$\beta < 1$$

La convergencia del método del gradiente depende del número de condición de A.





A, bien condicionada



A mal condicionada

Laboratorio:

$$\phi(x) = \frac{1}{2} x^T A x - b^T x$$

$$A_1 = \begin{pmatrix} 0.4343 & 0.4340 \\ 0.4340 & 0.4337 \end{pmatrix}, \quad b_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad x_1 = \frac{1}{9}\begin{pmatrix} -4337000 \\ 43400000 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 3 & -1 & 1 \\ -1 & 3 & -1 \\ 1 & -1 & 3 \end{pmatrix}, \quad b_2 = \begin{pmatrix} -1 \\ 7 \\ -7 \end{pmatrix}$$

$A_1 = >>> 1$

$A_2 > 1$

$$R_k = \frac{\phi(x_k) - \phi(x^*)}{\phi(x_{k-1}) - \phi(x^*)} < \beta < 1$$

$\phi \to A_1, b_1$

$\quad\; A_2, b_2 \quad < 1$

$$|\phi(x_k) - \phi(x^*)| = \beta |\phi(x_{-1}) - \phi(x^*)|$$

$$\underbrace{|\phi(x_k) - \phi(x^*)|}_{e_k} = \beta \underbrace{|\phi(x_{-1}) - \phi(x^*)|}_{e_{k-1}}$$

Valores y vectores propios

lunes, 8 de marzo de 2021 11:03

Dada una matriz $A \in \mathbb{C}^{n\times n}$, el problema de valores propios consiste en encontrar un escalar λ (real o complejo) y un vector no nulo x tq:

$$Ax = \lambda x,$$

donde λ es el valor propio y x su asociado vector propio. El vector propio no es único, pues todos sus múltiplos αx, con $\alpha \neq 0$ también son vectores propios asociados a λ.

Método de la potencia:

Consideramos matrices diagonalizables, i.e, matrices que se pueden descomponer de la sig. manera: $A = PDP^{-1}$.

Sean $\eta_1, \eta_2, \dots, \eta_n$ una base ortonormal de vectores propios de A.

$$A\eta_i = \lambda_i \eta_i, \quad i = 1, 2, \dots, n$$

y tales que $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$

$$x_0 = \sum_{i=1}^{n} \alpha_i \eta_i, \quad \alpha_1 = \langle x_0, \eta_1 \rangle \neq 0$$

$$= \langle (\alpha_1 \eta_1 + \alpha_2 \eta_2 + \dots + \alpha_n \eta_n), \eta_1 \rangle$$

$$\neq 0$$

Si x_0 no es ortonormal a y_1.

Ahora:

$$x^k = A^k x^0$$

$$= \sum_{i=1}^{n} \alpha_i A^k \eta_i$$

$$= \sum_{i=1}^{n} \alpha_i \lambda_i^k \eta_i$$

$$= \alpha_1 \lambda_1^k \eta_1 + \sum_{i=2}^{n} \alpha_i \lambda_i^k \eta_i$$

$$= \alpha_1 \lambda_1^k \left(\eta_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \frac{\lambda_i^k}{\lambda_1^k} \eta_i \right)$$

z_k

Recordemos $|\lambda_1| > |\lambda_i|$

$$\Rightarrow \quad (1)^k > \left| \frac{\lambda_i}{\lambda_1} \right|^k \geq \left(\frac{\lambda_i}{\lambda_1} \right)^k$$

Por lo tanto:

$$\lim_{k \to \infty} z_k = \lim_{k \to 0} \left(\eta_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \eta_i \right)$$

$$< 1$$

$$= \eta_1$$

como η_1 es normalizado, $\|z_k\| \to 1$.

Es decir $\|z_k\| \to 1$ y $z_k \to \eta_1$, por lo

Luego

$$\frac{x^k}{\|x^k\|} = \frac{\alpha_1 \lambda_1^k z_k}{|\alpha_1 \lambda_1^k| \, \|z_k\|}$$

$$= \frac{\operatorname{sign}(\alpha_1 \lambda_1^k) \, z_k}{\|z_k\|}$$

$$\Rightarrow \frac{x^k}{\|x^k\|} = \pm \frac{z_k}{\|z_k\|} \longrightarrow \pm v_1$$

vector propio

Método de la potencia

$x^{(0)}$ arbitrario y

$$y^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}$$

for $k = 1, 2, \ldots$

$$x^{(k)} = A y^{(k-1)}$$

$$y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}$$

$$\lambda^{(k)} = y^{(k)T} A y^{(k)}$$

end

$$\frac{x^T A x}{\|x\|^2} = \frac{x^T \lambda x}{\|x\|^2} = \lambda \frac{\|x\|^2}{\|x\|^2}$$

$$\lambda = \frac{x^T}{\|x\|} A \frac{x}{\|x\|}$$

Notemos que, por recursión, uno encuentra

$y^k = \beta^{(k)} A^k y^{(0)}$, con $\beta^{(k)} = \left(\prod_{i=1}^{k} \|x^{(i)}\| \right)^{-1}$ para

$k \geq 1$. Debido a la presencia de potencias de A se justifica el nombre del método.

de A se justifica el nombre del método.

El error $\|x^k - \lambda^1\|$ es proporcional a $\left|\frac{\lambda_2}{\lambda_1}\right|^k$.

Generalización del método de la potencia.

Una posible primera generalización del método de la potencia consiste en aplicarlo a la inversa de una matriz A (dando que A es no singular). Como los valores propios de A^{-1} son los recíprocos a los de A, el método de la potencia nos permite en este caso aproximar el valor propio de A de mínimo módulo.

Sea $x^{(0)}$ un vector arbitrario

$$y^{(0)} = \frac{x^{(0)}}{\|x^{(0)}\|}$$

for $k = 1, 2, \ldots$

$$x^{(k)} = A^{-1} y^{(k-1)} \quad (*)$$

$$y^{(k)} = \frac{x^{(k)}}{\|x^{(k)}\|}$$

$$\mu^{(k)} = y^{(k)T} A^{-1} y^{(k)}$$

Si A admite vectores propios li y si

los valores propios λ_4 de mínima norma son distintos de los otros, entonces

$$\lim_{k \to \infty} \mu^{(k)} = \frac{1}{\lambda_4}$$

i.e. $(\mu^k)^{-1} \longrightarrow \lambda_k$ para $k \to \infty$.

En cada paso (*) tenemos que resolver un sistema lineal $A(x^k) = y^{(k-1)}$! En este punto es conveniente efectuar una F.L.U de A (Cholesky si A es s.d.p) para resolver los sistemas lineales en cada iteración.

Desventajas:

① Sólo calculamos el vector propio asociado al valor propio más grande (en módulo).

② La velocidad de convergencia depende de $\left|\frac{\lambda_2}{\lambda_1}\right|$ si $\lambda_1 \approx \lambda_2$ el método converge muy lento ($\frac{|\lambda_{n+1}|}{|\lambda_n|}$ para el caso de la matriz inversa).

Nota: Supongamos que conocemos una estimación $\bar{\lambda} \approx \lambda_i$ de un valor propio arbitrario de A (diagonalizable) tq.

$$|\bar{\lambda} - \lambda_i| < |\bar{\lambda} - \lambda_j| \quad \forall j \neq i$$

entonces $(\bar{\lambda} - \lambda_i)^{-1}$ es el valor propio más grande en valor absoluto de la matriz $(A - \bar{\lambda} I)^{-1}$, así aplicamos el método de la potencia a la matriz $(A - \bar{\lambda} I)^{-1}$ obteniendo el esquema:

$$(A - \bar{\lambda} I) x^{k+1} = \frac{x^k}{\|x\|}, \quad k = 0, 1, \ldots$$

llamado el método inverso de la potencia.

(eig)

Criterio de parada:

$$\frac{|\lambda^k - \lambda^{k-1}|}{|\lambda^k|} \leq \text{tol}$$

$$A = \begin{pmatrix} \alpha & 2 & 3 & 13 \\ 5 & 11 & 10 & 8 \\ 9 & 7 & 6 & 12 \\ 4 & 14 & 15 & 1 \end{pmatrix}$$

$N = 37$

$\alpha = 30$

$\alpha = -30$

$x_2 = \alpha x$ (vect.)

$$Ax = \lambda x$$

x_2

$A(\alpha x) = Ax_2 = A(\alpha x)$

$= \alpha Ax$

$= \alpha \lambda x$

$= \lambda(\alpha x)$

x_2

Cálculo de todos los valores propios: QR

Dos matrices cuadradas A y B de la misma dimensión se denominan similares si existe una matriz no singular P tq.

$$P^{-1}AP = B$$

Las matrices similares comparten los mismos valores propios. De hecho, si λ es un valor propio de A y $x \neq 0$ es su vector propio asociado, tenemos:

$$BP^{-1}x = P^{-1}Ax \qquad (PAP^{-1} = B)$$

$$= P^{-1}\lambda x$$

$$= \lambda P^{-1}x.$$

Así λ es también valor propio de B ...

$= \lambda P^{-1}x$.

Así λ es también valor propio de B y su vector propio asociado es $y = P^{-1}x$.

Los métodos que admiten simultáneamente aproximaciones de todos los valores propios de una matriz se basan en la idea general de transformar a A (después de un número infinito de pasos) en una matriz similar que sea diagonal o triangular.

De esta manera, sus valores propios están dados en los entradas de la diagonal principal.

Entre los métodos principales tenemos el método QR.

QR

El método QR se denomina de esta manera ya que usa factorizaciones QR repetidamente para calcular los valores propios de A.

En este caso, vamos a estudiar el método sólo para matrices reales.

La idea consiste en construir una sucesión $A^{(k)}$, de matrices similares a A.

Se inicializa con $A^{[0]} = A$, y en cada

se inicializa con $A^{(0)} = A$, y en cada iteración k se usa la fact. QR para calcular matrices cuadradas $Q^{(k+1)}$ y $R^{(k+1)}$ tq.

$$Q^{(k+1)} R^{(k+1)} = A^{(k)}$$

y luego $\underline{A^{(k+1)} = R^{(k+1)} Q^{(k+1)}}$.

Las matrices $A^{(k)}$, $k = 0, 1, 2, \ldots$, son similares, por lo tanto, comparten los valores propios con A.

Si $A \in \mathbb{R}^{n \times n}$ y sus valores propios satisfacen $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$ entonces:

$$\lim_{k \to +\infty} A^{(k)} = T = \begin{bmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ & \lambda_2 & \ddots & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

En cada iteración k tenemos:

$$A^{(k)} = R^{(k)} Q^{(k)} = Q^{(k)^T} (Q^{(k)} R^{(k)}) Q^{(k)}$$

$$= Q^{(k)^T} A^{(k-1)} Q^{(k)}$$

$$= (Q^{(0)} Q^{(1)} \cdots Q^{(k)})^T A \underbrace{(Q^{(0)} Q^{(1)} \cdots Q^{(k)})}, \quad k \geq 0$$

$$= \underbrace{(Q^{(0)} Q^{(1)} \cdots Q^{(k)})}_{P^{-1}} \; A \; \underbrace{(Q^{(0)} Q^{(1)} \cdots Q^{(k)})}_{P}, \quad k \geq 0$$

Es decir, cada matriz $A^{(k)}$ es ortogonalmente similar a A.

Convergencia del método QR.

Sea $A \in \mathbb{R}^{n \times n}$ una matriz tq.

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$$

Entonces:

$$\lim_{k \to \infty} A^{(k)} = \begin{bmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ & \lambda_2 & \ddots & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Para la convergencia tenemos:

$$|t_{i,i-1}^{(k)}| = O\left(\left|\frac{\lambda_i}{\lambda_{i-1}}\right|^k\right), \quad i = 2, \ldots n, \text{ para } k \to +\infty$$

Bajo la suposición que A sea simétrica, la sucesión $\{A^{(k)}\}$ tiende a una matriz diagonal.

Nota: Si los valores propios de A no están "bien separados" entonces el método converge lentamente.

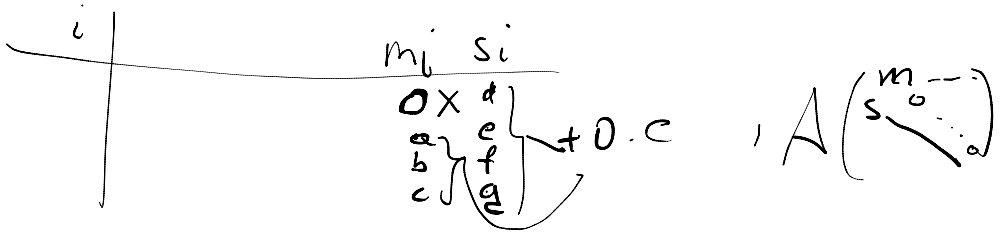
... acelerar las iteraciones

lentamente.

Nota 2: En la práctica, las iteraciones se paran cuando $\max_{i>j} |a_{i,d}^{(k)}| \leq tol$, $tol > 0$.

En matlab:

$\max(\max(abs(tril(A,-1)))) \leq tol$



c = último # cédula

$B_{Richar.} = I - A$

$B_{Jacobi} = (I - D^{-1}A) = -D^{-1}(L+U)$

$B_{GS} = -(D+L)^{-1}U = (I - (D+L)^{-1}A)$

Paulina López