FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

Pontificada Universidad Católica

MAT2605

Cálculo Científico I

Autor:
Sebastián Lepe V.

2 de julio de 2025

Índice

Motivación e Introducción

En el mundo al medir objetos, estos nos entrega un valor, sin embargo, por la naturaleza del objeto, este valor no es el real. Por lo que existe un error entre el valor obtenido al valor real.

Los errores pueden surgir por modelamiento, métodos númericos, computación, entre otros. Por ejemplo, queremos estudiar la ecuación diferencial $-\Delta u = f$, a veces necesitamos aproximar, generando errores de aproximación.

Estudiando la ecuación diferencial $-\Delta u = f$, determinando la función f, se forman errores f, en los coeficientes, en el dominio (generalmente se denota por Ω), condición de contorno, etc.

De forma reducida, cálculo científico I estudia métodos númericos con errores de aproximación.

1. Conceptos Básicos

Consideramos una máquina un instrumento limitado que puede representar número. Por ejemplo, la computadora, la calculadora, entres otros. El problema que la máquina solo puede mostrar números finitos. A cada máquina se le puede asociar los números de máquina (es un conjunto discreto).

Ejemplo: Consideremos la ecuación $x^2 + 62{,}10 + 1 = 0$. La solución son:

$$x_{1/2} = -\frac{62{,}10}{2} \pm \sqrt{\left(\frac{62{,}10}{2}\right)^2 - 1} = \{-62{,}08\ldots, 0{,}002\ldots\}$$

Tenemos un número central 31,05 y una parte irracional, si consideramos 4 dígitos, obtenemos,

$$-31{,}05 \pm 31{,}03 = \begin{cases} -0{,}02 \\ -62{,}08 \end{cases}$$

Ahora si trabajamos con 5 dígitos se genera un error con la mayor solución, en particular,

$$\frac{|-0{,}016 + 0{,}02|}{|-0{,}016|} \approx 25\,\%$$

Por otro lado,

$$\frac{|-62{,}08\cdots+62{,}08|}{|-62{,}08\ldots|} \leq \frac{0{,}0\ldots}{10} \leq \frac{0{,}01}{10} = 0{,}1$$

En conclusión,

(a) Hay, errores absolutos y relativos. Hay valores reales μ y aproximaciones $\widetilde{\mu}$. En particular, se define el absoluto por $|\mu - \widetilde{\mu}|$ y el relativo por $|\mu - \widetilde{\mu}|/|\mu|$.

(b) El cálculo de la raíz puede ser **intestable**. Decimos que es **estable** si al hacer pequeños cambios, entonces se generan pequeños cambios, en caso contrario decimos que es inestable. (Análogamente condicionado ¿?).

(c) Hay que evitar usar números casis iguales (evitar la cancelación de cifras/dígitos significativos).

Consideremos una máquina, definimos los números de la máquina por,

$$\mathcal{M} := \{\pm 0.a_1\ldots a_m 10^l, a_1 \neq 0, a_1, \ldots, a_m \in \{0, \ldots, 9\}, l \in \mathbb{Z}\,\text{acotado}\}$$

Decimos que están en la forma de punto flotante normalizado.

Existe un mapa de $x \in \mathbb{R} \mapsto fl(x) \in \widetilde{\mathcal{M}}$ (el conjunto $\widetilde{\mathcal{M}}$ es el conjunto $\mathcal{M}$ pero ignorando la cota de l.) Sea $x = 0.a_1\ldots a_m\ldots 10^l \in \mathbb{R}$, denotamos,

$$x' = 0.a_1\ldots a_m 10^l$$
$$x'' = 0.a_1\ldots(a_m + 1)10^l$$

Definimos el truncamiento por $fl(x) := x'$. Y definimos el redondeo $fl(x) := \text{round}(x) =$ número $\{x', x''\}$ más cercano a x, (estamos pensando con respecto al término m).

Ejemplo: Consideremos una máquina con números de máquina determinados por $m = 1$ y $-1 \leq l \leq 1$. Entonces,

$$\mathcal{M} = \{-9; -8; \ldots; -1; \ldots; 0; 0{,}01; 0{,}02; \ldots 0{,}1; 0{,}2; \ldots; 1; \ldots; 9\}$$

Encontremos una cota para el error relativo. Supongamos que l no tiene cota y tenemos m dígitos, es decir, estamos trabajando con un número de la forma $x = 0.a_1 \ldots a_m 10^l$. Consideremos el redondeo, por lo que,

$$x \mapsto fl(x) = \pm 0.a_1 \ldots \widetilde{a_m} 10^l$$

Luego se tiene que,

$$\begin{aligned} \frac{|x - fl(x)|}{|x|} &= \frac{|0.a_1 \ldots a_m \ldots 10^l - 0.a_1 \ldots \widetilde{a_m} 10^l|}{|x|} \\ &\leq \frac{|0{,}0 \ldots 5|10^l}{0{,}1 \cdot 10^l} = 5 \cdot 10^{-m} \end{aligned}$$

De esta forma se cunple que,

$$|x - fl(x)| \leq |x| 5 \cdot 10^{-m}$$

Definición: *Consideremos $x \in \mathbb{R}$ cualquiera, y sea x^* una aproximación de x (ya sea redondedo o truncamiento), definimos el* ***error absoluto*** *por,*

$$|x - x^*|$$

Y definimos el error relativo por,

$$\frac{|x - x^*|}{|x|}$$

El ejemplo anterior muestra cotas para el error absoluto y relativo.

Observación:

- Más adelante definiremos normas $\|\cdot\|$, pero la definición es simplemente análoga.
- El error relativo por redondeo es de orden 10^{-m} si la máquina usa m dígitos.
- Si $\text{round}(x) = 0.a_1 a_2 \ldots a_m 10^l$, se dice que conocemos x con m **cifras/dígitos significativos**.
- Del ejemplo anterior se puede concluir que $\text{round}(x) = x(1 + \delta)$ donde $|\delta| \leq \varepsilon = 5 \cdot 10^{-m}$. A ε se le llama **precisión** o **el ε de la máquina**. El verdadero ε se calcula con la base 2.

- Se cumple que,

$$|x - \text{round}(x)| \leq \varepsilon |x|$$

donde $\varepsilon = \text{mín}\{\delta > 0 : \text{round}(1+\delta) \neq 1\}$. Esto nos dice que ε es el menor delta tal que el redondeo de $1+\delta$ no sea 1.

Nota: ε no es el menor número que se puede representar en la máquina.

1.1. Operación Aritmética

En algunas ocasiones debemos operar las aproximaciones, por lo que definimos un operador que nos indique sumar, restar, multipicar y dividir usando aproximaciones.

Se define la operación,

$$\circledast \begin{cases} \mathbb{R} \times \mathbb{R} & \to \mathcal{M} \\ (x, y) & \mapsto x \circledast y = fl(fl(x) * fl(y)) \end{cases}$$

Donde $* \in \{+, -, \cdot, \}$.

Ejemplo: Consideremos los números,

$$x = \frac{5}{7}, \;\; y = \frac{1}{3}$$

con truncamiento de 5 dígitos significativos, entonces,

$$\begin{aligned} x \oplus y &= \text{fl}(\text{fl}(x) + \text{fl}(y)) \\ &= \text{fl}(0{,}71428 + 0{,}33333) \\ &= \text{fl}(1{,}04761) = 0{,}10476 \cdot 10^1 \end{aligned}$$

Esto último es así puesto que la máquina procesa solo 5 dígitos.

Observación:

- Restar números semejantes produce un error relativo grande. Y dividir por un número pequño o multiplicar por uno grande, aumenta el error absoluto pero no el relativo.
- Sumar números grandes y pequeños, puede producir un error absoluto grande pero no tanto error relativo.

Ejemplo: Sea la ecuación $x^2 + 62{,}10x + 1 = 0$. Pensemos en una máquina con mantisa $m = 4$ sin cota. La soluciones son,

$$x_{1,2} = -\frac{62{,}10}{2} \pm \sqrt{\left(\frac{62{,}10}{2}\right)^2 - 1}$$

Entonces aplicando la operación $\circledast$ para obtenemos números de máquina, se obtiene,

$$\begin{aligned} \widetilde{x}_1 &= -31{,}05 \oplus \sqrt{31{,}05 \otimes 31{,}05 \ominus 1} \\ &= -31{,}05 \oplus 31{,}03 \\ &= -0{,}02 \end{aligned}$$

1.2. Algoritmos y Estabilidad

Definición: *Un **algoritmo** es un procedimiento que describe, sin ambigüedades, una serie finita de instrucciones en un orden específico.*

Ejemplo: Dado los números reales $x_1, \ldots, x_N$, queremos calcular $S = \sum_{i=1}^{N} x_i$ mediante un algoritmo. Para ello necesitamos datos iniciales, por ejemplo, necesitamos N para saber cuántos números sumar y los N números $x_1, \ldots, x_N$. Luego se define suma= 0 y simplemente sumar suma con x_1, luego con x_2 y así hasta x_N. Finalmente el output es la suma total. Esto se describe de la siguiente forma:

input: $N, x_1, \ldots, x_N$ *(datos)*
(1) suma $= 0$
(2) para $i = 1, 2, \ldots, N$
suma $=$ suma $+ x_i$
output: suma

Definición: *Un algoritmo se llama **estable** si pequeños cambios en los datos producen pequeños cambios en los resultados. En caso contrario se llama **inestable**.*

Definición: *Sea $E_0 > 0$ el error inicial y sea E_n la magnitud del error después de n operaciones sucesivas. Si $E_n \approx C_n E_0$ con C una constante independiente de n, entonces se dice que el crecimiento del error es lineal. Por otro lado, si $E_n \approx C^n E_0$ con alguna constante $c > 01$, entonces el crecimiento se llama exponencial.*

Nota: Cuando hablamos de problemas, ya no decimos estabilidad sino condicionamiento del problema. Por lo que decimos que está bien condicionado si es estable y decimos que está mal condicionado si el número de condicionamiento es muy grande (generalmente se puede definir).

Nota: Puede haber un algoritmo inestable con problemas bien condicionados o algoritmos estables con problemas mal condicionados.

Ejemplo: Sea f una función derivable. Queremos calcular $f(x)$ con x real. Podemos perturbar x con un pequeño h, de forma que obtenemos $f(x+h)$. El error absoluto cumple lo siguiente,

$$|f(x+h) - f(x)| = |f'(\xi)h|$$

con ξ un valor entre $x+h$ y x. Para h pequeño se cumple que,

$$|f'(\xi)h| \approx |hf'(x)|$$

Si $|f'(x)|$ es pequeño, entonces $|f(x+h) - f(x)|$ es pequño y está bien condicionado.

Por otro lado podemos estudiar el error relativo,

$$\frac{|f(x+h) - f(x)|}{|f(x)|} \approx \frac{|hf'(x)|}{|f(x)|} = \frac{|xf'(x)|}{|f(x)|}\frac{|h|}{|x|}$$

Tomando $f(x) \neq 0, x \neq 0$. De aquí definimos el número de condición:

$$\mathcal{K}(x) := \left| \frac{xf'(x)}{f(x)} \right|$$

para $f(x) \neq 0$. Por ejemplo, si tomamos $f(x) = \arcsin(x)$, ocurre que está mal condicionada en $x = \pm 1$ al estudiar el número de condicionamiento.

1.3. Convergencia

Definición: *Una sucesión $\{\alpha_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}$ converge a α con orden $l \leq 1$ si y sólo si existen un $K > 0$ y $n_0 \in \mathbb{N}$ tales que,*

$$|\alpha_{n+1} - \alpha| \leq K|\alpha_n - \alpha|^l$$

Para todo $n \geq n_0$. Pediremos que $K < 1$ si $l = 1$. Diremos que la convergencia es lineal si $l = 1$, y diremos que es cuadrática si $l = 2$.

Definición: *Sean $\{\alpha_n\}_{n\in\mathbb{N}}, \{\beta_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}$ tales que $\beta_n > 0$ y sea $\alpha \in \mathbb{R}$. Diremos que,*

(a) $\alpha_n = \alpha + \mathcal{O}(\beta_n)$ para $n \to \infty$ si y sólo si existen $K > 0$ y $n_0 \in \mathbb{N}$ tales que,

$$|\alpha_n - \alpha| \leq K\beta_n$$

para todo $n \geq n_0$. (Si $\beta_n \to 0$, entonces α_n converge a α con orden/razón β_n).

(b) $\alpha_n = \alpha + o(\beta_n)$ para $n \to \infty$ si y sólo si,

$$\frac{|\alpha_n \alpha|}{\beta_n} \to 0$$

para $n \to \infty$.

Ejemplo: Sea $\alpha_n = 1/n$. Probemos que $\alpha_n = \mathcal{O}(1/n)$. Claramente,

$$|\alpha_n - 0| \leq \frac{K}{n} \Longleftrightarrow \frac{1}{n} \leq \frac{K}{n}$$

También se tiene que $\alpha_n = \mathcal{O}(1)$, puesto que,

$$|\alpha_n| \leq K$$

Y $\alpha_n = o(1)$, ya que,

$$\frac{|\alpha_n - 0|}{1} = \frac{1}{n} \to 0$$

para $n \to \infty$. Pensemos en $\alpha_n = 1/n^2$, entonces $\alpha_n = o(1/n)$, puesto que,

$$\frac{1}{n^2} / \frac{1}{n} = \frac{1}{n} \to 0$$

Para $n \to \infty$.

También podemos concluir que $\alpha_n = \mathcal{O}(1/n)$, pero además, $\alpha_n = \mathcal{O}(1/n^2)$ y esto dice mucho más o entre mas información que $\mathcal{O}(1/n)$.

Definición: *Sean F, G funciones tales que,*

$$\lim_{x \to x_0} G(x) = 0, \quad \lim_{x \to x_0} F(x) = L$$

Entonces,

(a) Si existe $K > 0$ tal que,

$$|F(x) - L| \leq K|G(x)|$$

para todo x en una vecindad de x_0, entonces decimos que $F(x) = L + \mathcal{O}(G(x))$ para $x \to x_0$.

(b) Sea $G(x) \neq 0$ en una vecindad de x_0 para $x \neq x_0$. Si,

$$\frac{|F(x) - L|}{|G(x)|} \to 0$$

cuando $x \to x_0$. Entonces escribimos $F(x) = L + o(G(x))$ para $x \to x_0$.

Ejemplo: Sea $\operatorname{sen}(x)$. Afirmamos que $\operatorname{sen}(x) = \mathcal{O}(x)$ cuando $x \to 0$, y en efecto, sabemos que para x muy pequeño, se cumple que $|\operatorname{sen}(x)| \leq |x|$, donde claramente $x \to 0$, y $\operatorname{sen}(x) \to 0$ cuando $x \to 0$.

Por otro lado se cumple que $\operatorname{sen}(x) = o(\sqrt{x})$, dado que $\sqrt{x} \neq 0$ en una vecindad de 0 sin incluir el 0, y se cumple que,

$$\frac{|\operatorname{sen}(x)|}{|\sqrt{x}|} \to 0$$

usando sandwich.

También podemos probar que $\cos(x) = 1 + \mathcal{O}(x^2)$ cuando $x \to 0$, puesto que,

$$|\cos(x) - 1| \leq e|x^2|$$

1.4. Eliminación de Gauss

En el siguiente capítulo estudiaremos sistemas lineales, por lo que es importante poder eleminar matrices usando números de máquinas

Ejemplo: Resolvamos la siguiente ecuación lineal utilizando cinco cifras.

$$\begin{pmatrix} 10^{-6} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Resolviendo la ecuación, se obtiene,

$$\begin{pmatrix} 10^{-6} & 1 \\ 0 & 1-10^{-6} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2-10^{-6} \end{pmatrix}$$

Entonces la solución exacta para y sería,

$$y = \frac{2-10^{-6}}{1-10^{-6}} = 0{,}9999989$$

Aplicando el punto flotante con un redondeo, se tiene que,

$$y^* = fl(y) = 1$$

y

$$x = 10^6(1-y) = \frac{1}{1-10^6}$$

Que no es cero, pero al aplicar punto flotante, se obtiene que $x^* = fl(x) = 0$. Por lo que la solución aproximada es $x^* = 0, y^* = 1$.

Observación: Este es un algoritmo inestable.

Ejemplo Consideremos un sistema lineal con tres cifras significativas.

$$\begin{pmatrix} 0-778 & 0{,}563 \\ 0{,}913 & 0{,}659 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0{,}215 \\ 0{,}254 \end{pmatrix}$$

La solución exacta es $x = 1, y = -1$. Pero usando número de máquina mediante redondeo se obtiene los resultados,

$$y^* = -0{,}5, \quad x^* = \frac{1}{0{,}913}(0{,}254 - 0{,}659y^*) = 0{,}640$$

El cual se aleja mucho de la solución exacta, por lo tanto se trata de un problema mal condicionado, en particular, la matriz es mal condicionada.

2. Solución de Sistemas Lineales

2.1. Métodos Directos: Eliminación de Gauss y Descomposición LU

Considermeos la ecuación lineal $ax = b$ con $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$. Podemos descomponer la matriz A en una matriz triangular superior.

Ejemplo: Consideremos la matriz,

$$\begin{pmatrix} 0{,}913 & 0{,}659 \\ 0{,}778 & 0{,}563 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0{,}254 \\ 0{,}215 \end{pmatrix}$$

Observemos que aplicando la iteración $(2) - (1)0{,}778/0{,}913$ en al segunda fila, obtenemos la matriz triangular superior,

$$\begin{pmatrix} 0{,}913 & 0{,}659 \\ 0 & 0{,}002 \end{pmatrix}$$

Expresado de forma más formal,

$$\begin{pmatrix} 1 & 0 \\ -\frac{0{,}778}{0{,}913} & 1 \end{pmatrix} \begin{pmatrix} 0{,}913 & 0{,}659 \\ 0{,}778 & 0{,}563 \end{pmatrix} = \begin{pmatrix} 0{,}913 & 0{,}659 \\ 0 & 0{,}002 \end{pmatrix}$$

$$\underbrace{\begin{pmatrix} 0{,}913 & 0{,}659 \\ 0{,}778 & 0{,}563 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 \\ -\frac{0{,}778}{0{,}913} & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 0{,}913 & 0{,}659 \\ 0 & 0{,}002 \end{pmatrix}}_{U}$$

Obteniendo la descomposición.

Estudiemos este proceso de forma general. Consideremos una matriz A de $n \times n$. Por el método de Gauss, es claro que podemos tomar la matriz A y transformala a una triangular superior, por lo que nos interesa saber los detalles.

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

Podemos tomar la matriz,

$$L_1 = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ m_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1} & 0 & \dots & 1 \end{pmatrix}$$

con $m_{i1} = -\frac{a_i 1}{a_{11}}$. Luego se observa que,

$$L_1 A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ 0 & a^1_{22} & \dots & a^2_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a^1_{N2} & \dots & a^1_{NN} \end{pmatrix}$$

Sea la matriz,

$$L = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & m_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & m_{N2} & 0 & \dots & 1 \end{pmatrix}$$

donde $m_{i2} = -\frac{a^1_{i2}}{a^1_{22}}$. Luego,

$$L_2L_1A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1N} \\ 0 & a^1_{22} & a^1_{23} & \dots & a^1_{2N} \\ 0 & 0 & a^2_{33} & \ddots & a^2_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a^2_{N3} & \dots & a^2_{NN} \end{pmatrix}$$

Después de $N-1$ apsos obtenems una matriz triangular superior U,

$$L_{N-1}L_{N-2}\dots L_2L_1A = U = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1N} \\ 0 & u_{22} & \dots & u_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & u_{NN} \end{pmatrix}$$

En particular $A = LU$ donde,

$$L := L_1^{-1}L_2^{-1}\dots L_{N-1}^{-1} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -m_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -m_{N1} & -m_{N2} & \dots & 1 \end{pmatrix}$$

Finalmente tenemos la descomposición $A = LU$

Teorema: *Sea $A \in \mathbb{R}^{N\times N}$ una matriz. Si no se intercambian las filas en la eliminación de Gauss, entonces existen U, L donde U es una triangular superior y L es triangular inferior con 1 en la diagonal tales que,*

$$A = LU$$

Definición: *La representación $A = LU$ se llama* ***descomposición LU*** *de A.*

Observación:

- Utilizar sólo A para almacenar L y U.
- En el caso con cambio de filas, se tiene,

  $$PA = LU$$

  donde P es una matriz de permutación P.
- **Pivoteo parcial** es por cambios de filas, **pivoteo total** es por cambios de filas y/o columnas. Cambiar las columnas de A requiere de otra permutación a la derecha de la igualdad.
- Supongamos que no hubo ninguna permutación, entonces,

  $$\det(A) = \det(U) = \prod_{i=1} Nu_{ii}$$

Ejemplo:

Este procedimiento, de resolver un sistema lineal con matriz triangular superior, se llama **sustitución regresiva**.

Ejemplo: Consideremos la misma matriz anterior,

$$\begin{pmatrix} 2 & 3 & 1 \\ 0 & 1 & 3 \\ 3 & 2 & 1 \end{pmatrix}$$

Ejemplo...

El costo computacional: Se mide el trabajo/tiempo necesario para contar el númer de **operaciones esenciales**, estas son multiplicaciones/divisiones y sumar/restas.

El costo de la eliminación de Gauss: Consideremos es sistema lineal $Ax = b$ donde $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ y $b = (b_i) \in \mathbb{R}^N$, entonces para la resolución de x, se estudia la matriz ampliada $(A|b) = (a_{ij}) \in \mathbb{R}^{N\times(N+1)}$. Generando un costo de eliminación.

- **Suma/Resta:** El costo de eliminación es,

$$\sum_{i=1}^{N-1}(N-i)(N-i+1)$$

- **Multiplicación/División:**

$$\sum_{i=1}^{N-1}(N-i)(N-i+2)$$

También se genera un costo en la sustitución.

- **Suma/Resta:**

$$\sum_{i=1}^{N-1}(N-i)$$

- **Multipliación/División:**

$$\sum_{i=1}^{N-1}(N-i+1)+1$$

Si resolvemos las sumatorias, obtenemos la siguiente tabla, Por tanto la eliminación de Gauss es un algoritmo caro.

	Mult/Div	Sum/Rest
Elim	$\frac{N^3}{3} + \frac{N^2}{2} - \frac{5}{6}N$	$\frac{N^3}{3} - \frac{N}{3}$
Sust	$\frac{N^2}{2} + \frac{N}{2}$	$\frac{N^2}{2} - \frac{N}{2}$
Total	$\frac{N^3}{3} + N^2 - \frac{1}{3}N$	$\frac{N^3}{3} + \frac{N^2}{2} - \frac{5N}{6}$

2.2. Descomposición de Cholesky

Definición: *Sea $A \in \mathbb{R}^{N\times N}$ una matriz. Decimos que es **definida positiva** si es simétrica y,*

$$x^T Ax > 0$$

para todo $x \in \mathbb{R}^N \setminus \{0\}$.

Nota: La definición general dice dada una matriz $A \in \mathbb{C}^{N\times N}$ es definida positiva si es hermitiana $A = (A^T)^*$ y para todo x no nulo, se tiene que $x^T Ax > 0$. Como A son matrices reales, se tiene que $A = A^T$.

Observación:

- Una matriz puede cumplir la propiedad definida positiva sin ser simétrica.
- **Teorema:** *Para $A = (a_{ij})_{i,j=1}^N$, las matrices $A_k := (a_{ij})_{i,j=1}^k$ con $(1 \leq k \leq N)$ se llaman submatrices principales. Entonces es definida positiva si y sólo si para todo $k = 1, \ldots, N$, se tiene que* $\det(A_k) > 0$.
- **Teorema:** *Una matriz simétrica es definida positiva si y sólo todos sus valores propios son positivos.*
- Sea A definida positiva, entonces la eliminación de Gauss funciona sin cambios de filas/-columnas y es estable.

Ejemplo:

1. Consideremos la matriz,

$$\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$$

   Entonces es definida positiva, puesto que,

$$\begin{aligned} \det(A_1) &= 2 > 0 \\ \det(A_2 = A) &= 5 > 0 \end{aligned}$$

2. Consideremos la matriz,

$$\begin{pmatrix} -2 & 1 \\ 1 & -3 \end{pmatrix}$$

   No es definida positiva, puesto que $\det(A_1) = -2$.

3. Consideremos la matriz,

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

Es definida positiva puesto que,

$$\begin{aligned} \det(A_1) &= 2 > 0 \\ \det(A_2) &= 7 > 0 \\ \det(A_3 = A) &= 10 > 0 \end{aligned}$$

El problema es que no se puede decidir si una matriz grande es definida positva, por la gran cantidad de términos, esto dio origen a la descomposición de Cholesky.

Teorema (Descomposición de Cholesky): *Sea $A \in \mathbb{R}^{N\times N}$ una matriz. Entonces A es definida positiva si y sólo si L una matriz triangular inferior con elementos positivos y diagonal no nula, tal que $A = LL^T$.*

Demostración: Supongamos que existe tal matriz inferior con coeficientes positivos tal que $A = LL^T$. Luego para todo $x \in \mathbb{R}^{N\times N}$,

$$\begin{aligned} x^T Ax &= x^T LL^T x \\ &= (L^T x)^T L^T x \\ &= \|L^T x\|_2^2 \end{aligned}$$

Supongamos que $L^T x = 0$, entonces necesariamente $x = 0$, cosa imposible, por tanto $L^T x \neq 0$ para todo $x \in \mathbb{R}^N$, de forma que $\|L^T x\|_2 > 0$, de forma que,

$$x^T Ax > 0$$

para todo $x \in \mathbb{R}^N$ no nulo.

Para la otra dirección es simplemente por inducción **por hacer apuntes cap 2**. ■

Ejemplo: Consideremos la siguiente matriz definida positiva,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix}$$

De esta forma,

$$\begin{aligned} l_{11} &= \sqrt{a_{11}} \\ l_{21} &= \frac{a_{12}}{l_{11}} \\ l_{31} &= \frac{a_{13}}{l_{11}} \\ l_{22} &= \sqrt{a_{22} - l_{21}^2} \\ l_{32} &= \frac{a_{23} - l_{21}l_{31}}{l_{22}} \\ l_{33} &= \sqrt{a_{33} - l_{31}^2 - l_{32}^2} \end{aligned}$$

Encontrando una fórmula para matrices 3×3.

En el caso general $A \in \mathbb{R}^{N\times N}$ se tiene lo siguiente. Si $A = LL^T$, entonces,

$$a_{ik} = \sum_{j=1}^{k} l_{ij}l_{kj} = \sum_{j=1}^{\text{mín}\{i,k\}} l_{ij}l_{kj}$$

para todo $1 \leq i \leq k \leq N$.

Ejemplo:

-
- Consideremos la siguiente matriz,

$$\begin{pmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 14 \end{pmatrix}$$

  Determinemos la descomposición de Cholesky. Para ello debemos ver que es definida positiva, cosa que se cumple puesto que $\det(A_k) > 0$ para todo $k = 1, 2, 3$. De la fórmula anterior podemos comprobar que la matriz L es,

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 2 & 1 & 3 \end{pmatrix}$$

2.3. Normas Vectoriales

Como estamos trabajando con espacios vectoriales de dimensión finita, estos son isomorfo a $\mathbb{R}^m$ para algún m, por lo que simplemente trabajaremos directamente con $\mathbb{R}^n$. Ahora queremos medir errores, y para ello necesitamos normas.

Definición: *Una norma es una función* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ *si cumple la siguientes condiciones:*

(i) $\|x\| \geq 0$ *para todo* $x \in \mathbb{R}^n$ *y* $\|x\| = 0$ *si y sólo si* $x = 0$.

(ii) $\|\alpha x\| = |\alpha|\|x\|$ *para todo* $x \in \mathbb{R}^n, \alpha \in \mathbb{R}$.

(iii) $\|x + y\| \leq \|x\| + \|y\|$ *para todo* $x, y \in \mathbb{R}^n$.

Ejemplo: Algunos ejemplos de norma son,

- $\|x\|_2 = \sqrt{x \cdot x}$ para todo $x \in \mathbb{R}^n$.
- $\|x\|_\infty = \text{máx}_{i=1,\ldots,n} |x_i|$.
- $\|x\|_1 = \sum_{i=1}^n |x_i|$.
- $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ (esto constituyen un subespacio de l_p con $p \geq 1$.)

Probemos que son normas.

- La primera y segunda condición son evidente. Para la tercera basta notar que,

$$(x + y) \cdot (x + y) = x \cdot x + 2x \cdot y + y \cdot y$$
$$es$$

- ....

temrinar

Desigualdad de Cauchy-Schawrz: *Para todo* $x, y \in \mathbb{R}^n$ *se cumple que,*

$$x^T y \leq \|x\|_2 \|y\|_2$$

Aplicaciones de la norma: Algunas aplicaciones para las normas son:

- Medir distancia entre vectores,

$$\text{dist}(x, y) = \|x - y\|$$

  para todo $x, y \in \mathbb{R}^N$.

- Convergencia de sucesiones de vectories: Sea $\{x_k\}_{k \in \mathbb{N}}$ una sucesión que converge a $x \in \mathbb{R}^N$. Entonces escribimos $x_k \to x$ cuando $k \to \infty$ si y sólo si $\|x_k - x\| \to 0$ cuando $k \to infty$. Por ejemplo,

$$x_k := \begin{pmatrix} 1 \\ 2 + \frac{1}{k} \\ \frac{1}{k^2} \\ e^{-k} \operatorname{sen}(k) \end{pmatrix} \to \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix}$$

  cuando $k \to \infty$.

Observación: La norma euclidiana es puesto que la podemos relacionar con el producto interior:

$$\|x\|_2^2 = \sum_{i=1}^N x_i^2 = x^T \cdot x$$

Observemos lo que ocurre en $\mathbb{R}^N$. Sabemos que $\|\cdot\|_p$ es una norma bien definida para $1 \leq p \leq \infty$. Si tomamos $p = 1, 2$ obtenemos las normas $\|\cdot\|_1, \|\cdot\|_2$, pero aquí ocurre lo interesante. Sea $x \in \mathbb{R}^n$, entonces,

$$\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \leq |x_1| + \cdots + |x_n| = \|x\|_1$$

Y también se cumple que,

$$\begin{aligned}\|x\|_1 &= |x_1| + \cdots + |x_n| \\ &= x_1\text{sgn}(x_1) + \cdots + x_n\text{sgn}(x_n) \\ &= x^T\text{sgn}(x) \\ &\leq \|x\|_2\|\text{sgn}(x)\|_2 = \sqrt{N}\|x\|_2\end{aligned}$$

Por lo tanto, para todo $x \in \mathbb{R}^N$ se cumple que,

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{N}\|x\|_2$$

Dicho de otra forma, las normas $\|\cdot\|_1, \|\cdot\|_2$ son por así decirlo "parecido".

Definición: *Sean $\|\cdot\|_a, \|\cdot\|_b$ dos normas en un espacio vectorial V. Decimos que son equivalentes si existen $C_1, C_2 > 0$ tales que para todo $x \in V$ se cumple que,*

$$C_1\|x\|_a \leq \|x\|_b \leq C_2\|x\|_a$$

Teorema: *Todas las normas en un espacio vectorial finito dimensional son equivalentes.*

Observación: Normalmente c_1, c_2 depende de n (dimensión del espacio).

2.4. Condicionamiento de Sistemas Lineales y Normas Matriciales

Consideremos el siguiente sistema lineal $Ax = b$. Podemos perturbar los datos de la siguiente forma tomando $b + \delta b$, generando $x + \delta x$, de forma que,

$$A(x + \delta x) \Longleftrightarrow A\delta x = \delta b$$

Nos interesa estudiar la relación,

$$\frac{\|\delta x\|}{\|x\|}, \quad \frac{\|\delta b\|}{\|b\|}$$

Veamos el siguiente caso. Sea $A \in \mathbb{R}^N$ una matriz general invertible. Tenemos,

$$\frac{\|Ax\|}{\|x\|}$$

Nos interesa encontrar el valor máximo de la razón, definimos,

$$\|A\| := \max_{x \in \mathbb{R}^N \setminus \{0\}} \frac{\|Ax\|}{\|x\|}$$

Definiendo una especie de norma para una matriz. También se cumple que,

$$\begin{aligned}\min_{x\in\mathbb{R}^N\setminus 0}\frac{\|Ax\|}{\|x\|} &= \min_{y=A^{-1}x, x\in\mathbb{R}^N\setminus\{0\}}\frac{\|Ay\|}{\|y\|}\\ &= \min_{x\in\mathbb{R}^N\setminus\{0\}}\frac{\|x\|}{\|A^{-1}x\|}\\ &= \frac{1}{\max_{x\in\mathbb{R}^N\setminus\{0\}}\frac{\|A^{-1}x\|}{\|x\|}}\\ &= \frac{1}{\|A^{-1}\|}\end{aligned}$$

Ahora observemos que dada una pertubación $a\delta x = \delta b$, se tiene que,

$$\begin{aligned}\|\delta x\| = \|A^{-1}\delta b\| &\leq \max_{x\in\mathbb{R}^N\setminus\{0\}}\frac{\|A^{-1}x\|}{\|x\|}\|\delta b\|\\ &= \|A^{-1}\|\|\delta b\|\end{aligned}$$

Por tanto,

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

Consideremos $\|\cdot\|$ un norma vectorial. Sea el sistema lineal $Ax = b$ con b no nulo y A regular (invertible). Se tiene que,

$$\|b\| = \|Ax\| = \frac{\|Ax\|}{\|x\|}\cdot\|x\| \leq \|A\|\cdot\|x\|$$

Entonces,

$$\frac{\|Ax\|}{\|x\|} \leq \max_{y\neq 0}\frac{\|Ay\|}{\|y\|}$$

Siempre y cuando $x \neq 0$. Definimos,

$$\|A\| := \max_{y\neq 0}\frac{\|Ay\|}{\|y\|}$$

Donde la notación de norma no es casualidad. A partir de una norma vectorial $\|\cdot\|$ podemos definir una especie de normal matricial. En particular, $\|A\|$ satisface las siguientes propiedades:

- $\|A\| \geq 0$, $\|A\| = 0$ si y sólo si $A = 0$.
- Si $\alpha \in \mathbb{R}$, entonces $\|\alpha A\| = |\alpha|\|A\|$.
- $\|A + B\| \leq \|A\| + \|B\|$
- $\|AB\| \leq \|A\|\|B\|$

Esto motiva la norma matricial.

Definición: *Sea una aplicación* $\|\cdot\| : \mathbb{R}^{N\times N} \to \mathbb{R}$ *que satisface las cuatro propiedades anteriores, entonces le llamamos norma matricial.*

Observación: La norma definida por la forma,

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

Se le llama norma matricial inducida por una norma vectorial $\|\cdot\|$ (abuso de notación).

Propiedades: Se cumple las siguientes propiedades de normal matricial inducida y vectorial:

- $\|A\| = \max_{x\in\mathbb{R}^N, \|x\|=1} \|Ax\|$.
- $\|Ax\| \leq \|A\|\|x\|$ para todo $A \in \mathbb{R}^{N\times N}$ y para todo $x \in \mathbb{R}^N$.
- $\|I\| = 1$ (normal matricial).
- $\|A \cdot B\| \leq \|A\|\|B\|$ para todo $A, B \in \mathbb{R}^{N\times N}$.

Observación: Podemos medir la distancia entre matrices. Simplemente se define por $\text{dist}(A, B) := \|A - B\|$. De esta forma podemos decir que una matriz A_k converge a A para $k \to \infty$ si y sólo si $\|A_k - A\| \to 0$ cuando $k \to \infty$.

Definición: *Sea* $\|\cdot\|$ *una norma matricial y* $A \in \mathbb{R}^{N\times N}$*, se define,*

$$cond(A) = \|A\|\|A^{-1}\|$$

con A invertible.

Propiedades: Sea $A \in \mathbb{R}^{N\times N}$ una matriz invertible. Se cumple las siguientes propiedades:

- Decimos que A es mal condicionada si $\text{cond}(A)$ es grande. En otro caso, decimos que es bien condicionada.
- El número de condición de una matriz depende de la norma.
- $\text{cond}(A) = \text{cond}(A^{-1})$.
- $\text{cond}(A) \geq 1$ para todo $A \in \mathbb{R}^{N\times N}$.
- $\text{cond}_2(A) = |\lambda_{\text{máx}}(A)|/|\lambda_{\text{mín}}(A)|$ si $A = A^T$.

Teorema: *Sea* $Ax = b$ $A \in \mathbb{R}^{N\times N}$ *regular,* $\delta b \in \mathbb{R}^N$ *tal que,*

$$A(x + \delta x) = b + \delta b$$

Entonces,

$$\frac{\|\delta x\|}{\|x\|} \leq cond(A)\frac{\|\delta b\|}{\|b\|}$$
$$cond(A)^{-1}\frac{\|\Delta b\|}{\|b\|} \leq \frac{\|\Delta x\|}{\|x\|}$$

Demostración: La primera parte ya fue vista. Para la segunda parte, basta en pensar que,

$$A(x+\Delta x) = b + \Delta b \longleftrightarrow A^{-1}(b+\Delta b) = x + \Delta x$$

Por tanto,

$$\text{cond}(A)^{-1}\frac{\|\Delta b\|}{\|b\|} \leq \frac{\|\Delta x\|}{\|x\|}$$

■

Teorema: *Sean* $A \in \mathbb{R}^{N\times N}, \Delta A \in \mathbb{R}^{N\times N} b \in \mathbb{R}^N$ *tales que* $Ax = b$ *y* $(A+\Delta A)(x+\Delta x) = b$ *. Entonces para normas compatibles se cumple que,*

$$\frac{\|\Delta x\|}{\|x+\Delta x\|} \leq cond(A)\frac{\|\Delta A\|}{\|A\|}$$

Y si además $cond(A)\frac{\|\Delta A\|}{\|A\|} \leq \delta < 1$*, entonces,*

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{cond(A)}{1-\delta}\frac{\|\Delta A\|}{\|A\|}$$

Demostración: Si $Ax = b$, entonces,

$$(A+\Delta A)(x+\Delta x) = b \Longleftrightarrow \Delta x = -A^{-1}\Delta A(x+\Delta x)$$

Entonces,

$$\frac{\|\Delta x\|}{\|x+\Delta x\|} = \frac{\|A^{-1}\Delta A(x+\Delta x)\|}{\|x+\Delta x\|} \leq \|A^{-1}\Delta A\| \leq \text{cond}(A)\frac{\|\Delta A\|}{\|A\|}$$

Por lo tanto,

$$\frac{\|\Delta x\|}{\|x\|} \leq \text{cond}(A)\frac{\|\Delta A\|}{\|A\|}$$

Ahora supongamos que $\text{cond}(A)\frac{\|A\|}{\|A\|} \leq \delta < 1$, entonces,

$$\frac{\|\Delta x\|}{\|x\|} = \frac{\|A^{-1}A(x+\Delta x)\|}{\|x\|} \leq \text{cond}(A)\frac{\|\Delta A\|}{\|A\|}\frac{\|x+\Delta x\|}{\|x\|}$$

Por otro lado se cumple que,

$$\frac{\|x+\Delta x\|}{\|x\|} \leq \frac{1}{1-\text{cond(A)}\frac{\|\Delta A\|}{\|A\|}} \leq \frac{1}{1-\delta}$$

Por la hipótesis del enunciado, finalmente se concluye qyuem

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|\Delta x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1-\delta}\frac{\|\Delta A\|}{\|A\|} \tag{1}$$

Probando el teorema. ■

Ejemplo:

- Consideremos las siguientes matrices,

$$A = \begin{pmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

con $\varepsilon < 1$ muy pequeño. Observemos que,

$$A^{-1} = \varepsilon^{-2} \begin{pmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{pmatrix}$$

Entonces $\|A\|_\infty = 2+\varepsilon$ y $\|A^{-1}\|_\infty = \frac{2+\varepsilon}{\varepsilon^2}$. De forma que,

$$\begin{aligned} \text{cond}_\infty(A) &= \left(\frac{2+\varepsilon}{\varepsilon}\right)^2 > \frac{4}{\varepsilon^2} = \mathcal{O}(\varepsilon^{-2}), \quad (\varepsilon \to 0) \\ \frac{\|A-B\|_\infty}{\|A\|_\infty} &= \frac{\varepsilon}{2+\varepsilon} < \frac{\varepsilon}{2} \end{aligned}$$

Se tiene que A está muy mal condiconada y que aparte, está muy cerca relativamente de una matriz singular. Esto es el significado de estar mal condicionadamente. Observemos que esto no tiene nada que ver con que su determinante sea pequeño, puesto que,

$$\det(A) = \varepsilon^2$$

Los siguientes dos ejemplo muestra este hecho.

- Consideremos la siguiente matriz,

$$A = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix} = \varepsilon \cdot I$$

Luego el determinante es $\det(A) = \varepsilon^2$ y la condicionalidad con respecto a cualquier matriz matricial inducida es,

$$\text{cond}(A) = \|A\|\|A^{-1}\| = \|I\|\|I^{-1}\| = 1$$

Siempre y cuando $\varepsilon \neq 0$.

- Consideremos la matriz,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon^{-1} \end{pmatrix}$$

Entonces el determinante es $\det(A) = \varepsilon^{-1}$. Con respecto a la condicionalidad se cumple que,

$$\text{cond}_\infty(A) = \text{cond}_1(A) = \text{cond}_2(A) = \frac{1}{\varepsilon}$$

Definición: *Sea A una matriz cuadrada, definimos el radio espectral de A por,*

$$\rho(A) := \text{máx}\{|\lambda| : \lambda \text{ es valor propio de } A\}$$

Teorema: *Consideremos $A = (a_{ij}) \in \mathbb{R}^{N\times N}$. Para $1 \leq p \leq \infty$, definimos*

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

Entonces se cumple que,

$$\|A\|_\infty = \max_{1\leq i\leq N} \sum_{j=1}^{N} |a_{ij}|$$
$$\|A\|_1 = \max_{1\leq i\leq N} \sum_{j=1}^{N} |a_{ij}|$$
$$\|A\|_2 = \sqrt{\rho(A^T A)}$$

Ejemplo: Consideremos la siguiente matriz,

$$A = \begin{pmatrix} 1 & 2 & -3 \\ 1 & -1 & 1 \\ 1 & 2 & 2 \end{pmatrix}$$

Entonces,

$$\|A\|_1 = \text{máx}\{3, 5, 6\} = 6$$
$$\|A\|_\infty = \text{máx}\{6, 3, 5\} = 6$$

Observemos en particular que $\|Ax\|_\infty \leq 6\|x\|_\infty$ para todo $x \in \mathbb{R}^3$. Estos nos dice que $\|Ax\|_\infty/\|x\|_\infty \leq 6$ es acotado y está bien definido.

Ejemplo: Sea $A = A^T \in \mathbb{R}^{N\times N}$. Afirmamos que A es diagonizable, en particular existe Q ortogonal ($Q^T Q = I$) tal que,

$$A = Q^T \text{diag}(\lambda_1, \ldots, \lambda_n) Q$$

donde $\lambda_1, \ldots, \lambda_n$ son los valores propios de A. Luego con respecto a la norma vectorial 2 sobre la matriz Q se cumple que,

$$\begin{aligned} \|Qx\|_2^2 &= (Qx)^T (Qx) \\ &= x^T Q^T qx = x^T x = \|x\|_2^2 \end{aligned}$$

Es decir, Q es isometría con respecto a la norma vectorial $\|\cdot\|_2$. Ahora determinemos la norma matricial inducida de A, pero antes debemos observar lo siguiente,

$$\begin{aligned}\|A\|_2 &= \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_\infty} \\ &= \max_{x\neq 0} \left\| \frac{Ax}{\|x\|_2} \right\| \\ &= \max_{x\neq 0} \left\| A\left(\frac{x}{\|x\|_2}\right) \right\|\end{aligned}$$

Observemos que la norma de $x/\|x\|_\infty$, entonces podemos afirmar que,

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$$

En particular no usamos ninguna condición de la norma matricial 2, por lo que esto se puede generalizara toda normal matricial inducida. Por tanto,

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

Ahora tenemos que,

$$\begin{aligned}\|A\|_2 &= \max_{\|x\|_2=1} \|Ax\|_2 \\ &= \max_{\|x\|_2=1} \|Q^T DQx\|_2\end{aligned}$$

donde $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Luego, si Q es isometría, entonces Q^T también lo es, de forma que,

$$\|A\|_2 = \max_{\|x\|_2=1} \|DQx\|_2$$

Tomando $Qx = y$, se tiene que $\|y\|_2 = \|Qx\|_2 = \|x\|_2 = 1$, de forma que,

$$\|A\|_2 = \max_{\|y\|_2=|} \|Dy\| = \max\{|\lambda_1|, \ldots, |\lambda_n|\} = \rho(A)$$

Finalmente se concluye que si $A = A^T$, entonces

$$\|A\|_2 = \rho(A)$$

Ejemplo: Sea $A \in \mathbb{R}^{N\times N}$ cualquiera. Queremos acotar $\|A\|_2$, para ello notemos que,

$$\begin{aligned}\|Ax\|_2^2 &= x^T A^T Ax \\ &= (x^T)(A^T Ax) \\ &\leq \|x^T\|_2 \|A^T Ax\|_2 \\ &\leq \|x^T\|_2 \|A^T A\|_2 \|x\|_2\end{aligned}$$

Si consideramos $\|x\|_2 = 1$, entonces

$$\|Ax\|_2^2 \leq \|A^T A\|_2$$

Si $A^T A$ es simétrico, entonces,

$$\|Ax\|_2^2 \leq \rho(A^T A)$$

Finalmente $\|Ax\|_2 \leq \rho(A^T A)$.

Ejemplo: Probemos la propiedad de condicional $\text{cond}_2(A)$. Observemos que,

$$\text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 = \rho(A)\rho(A^{-1})$$

Sabemos que $\rho(A) = |\lambda_{\text{máx}}|$, veamos que pasa con $\rho(A^{-1})$. Sea λ valor propio de A, entones existe x no nulo tal que,

$$Ax = \lambda x$$

Usando que A es invertible esto implica no tiene valor propio 0, luego se tiene que,

$$A^{-1}x = \frac{1}{\lambda}x$$

Es decir, $1/\lambda$ es valor propio de A^{-1} por lo que,

$$\rho(A^{-1}) = \text{máx}\{|1/\lambda_1|, \ldots, |1/\lambda_n|\}$$

Por tanto $\rho(A^{-1}) = 1/|\lambda_{\text{mín}}|$. Finalmente,

$$\text{cond}_2(A) = \rho(A)\rho(A^{-1}) = \frac{|\lambda_{\text{máx}}|}{|\lambda_{\text{mín}}|}$$

Demostración de las normas:

- **Norma $\|\cdot\|_2$:** Por lo que hemos visto evidentemente,

$$\|A\|_2^2 \leq \rho(A^T A)$$

- **Norma $\|\cdot\|_1$:**
- **Norma $\|\cdot\|_\infty$:**

Observación: Sea $A \in \mathbb{R}^{N\times N}$. Se define la norma Frobenius por,

$$\|A\|_F = \left(\sum_{i,j=1,\ldots,n} |a_{ij}|^2 \right)^{1/2}$$

Es importante notar que $\|A\|_2 \neq \|A\|_F$.

¿Habrá una relación entre la norma matricial $\|A\|$ y su radio espectral $\rho(A)$? El siguiente resultado verifica este hecho.

Teorema: *Sea $A \in \mathbb{R}^{N\times N}$. Entonces para toda norma inducida $\|\cdot\|$ se cumple que,*

$$\rho(A) \leq \|A\|$$

Demostración: Sea λ un valor propio de A, por lo que $Ax = \lambda x$ para un vector propio x no nulo. Entonces,

$$\|Ay\| = |\lambda|\|y\|$$

Por otro lado $\|Ay\| \leq \|A\|\|y\|$, entonces,

$$|\lambda| \leq \|A\|$$

Esto se puede hacer para todo valor propio de A, por tanto,

$$\rho(A) \leq \|A\|$$

■

Nota: Se puede demostrar que,

$$\rho(A) = \text{ínf}\{\|A\| : \ \|\cdot\| \text{es la norma matricial}\}$$

Y es claramente por definición

Proposición: *Sea $A \in \mathbb{R}^{N\times N}$ una matriz. Entonces las siguientes afirmaciones son equivalentes,*

(i) $A^k \to 0$ *para* $k \to \infty$.

(ii) $\|A^k\| \to 0$ *para* $k \to \infty$ *para alguna norma matricial inducida.*

(iii) $\rho(A) < 1$.

(iv) $A^k x \to 0$ *para* $k \to \infty$ *para todo* $x \in \mathbb{R}^N$.

Demostración:

- **(i) implica (ii):** Supongamos que $A^k \to 0$ cuando $k \to \infty$. Entonces,

$$\|A^k\|_\infty = \max_{1\leq i\leq N} \sum_{j=1}^{N} |a_{ij}^k| \to 0$$

  para $k \to \infty$.

- **(ii) implica (iii):** Sea λ valor propio de A, entonces λ^k es valor propio de A^k, luego,

$$\rho(A)^k \leq \rho(A^k) \leq \|A^k\| \overset{k\to\infty}{\longrightarrow} 0$$

  Si $\rho(A) \geq 1$, entonces $\rho(A)^k \not\to 0$, siendo contradicción, por lo tanto $\rho(A) < 1$.

- **(iii) implica (iv):** Sea $\varepsilon > 0$ tal que $\rho(A) + \varepsilon < 1$, entonces existe $\|\cdot\|$ norma inducida tal que $\|A\| < \rho(A) + \varepsilon$, entonces,

$$\|A^k x\| \leq \|A^k\|\|x\| < (\rho(A+\varepsilon)^k \|x\| \overset{k\to\infty}{\longrightarrow} 0$$

  Esto para todo $x \in \mathbb{R}^N$.

- **(iv) implica (i):** Basta tomar convenientemente x, en particular se toma $x = e_i$ para $i = 1, \ldots, N$. Esto implica que cada columna de A converge a 0, por tanto toda la matriz A converge a 0.

Demostrando la proposición. ■

2.5. Métodos Iterativos

Consideremos el sistema lineal $Ax = b$. Queremos descomponer A y obtener la siguiente expresión $x = Tx + c$ donde C es una transformación y c una constante. También queremos generar una secuencia $\{x^{(k)}\}_{k\in\mathbb{N}}$ definida de la siguiente forma $x^{(k+1)} = Tx^{(k)} + c$ de tal forma que $x^{(k)} \to x$ para cuando $k \to \infty$.

Jacobi: El primer método que veremos es el de Jacobi, queremos descomponer la matriz A de la siguiente forma,

$$A = -L + D - U$$

dode $-L$ es una triangular inferior, D la diagonal de A y $-U$ la triangular superior.

Si D es invertible, entonces,

$$\begin{aligned} Ax = b &\iff (-L + D - U)x = b \\ &\iff Dx = (L+U)x + b \\ &\iff x = D^{-1}(L+U)x + D^{-1}b \end{aligned}$$

De esta forma obtenemos $x = Tx + c$ donde $T = D^{-1}(L+U)$ y $c = D^{-1}b$. Finalmente definimos la iteración de Jacobi por:

$$x^{(k+1)} = D^{-1}(L+U)x^{(k)} + D^{-1}b$$

para todo $k = 0, 1 \ldots$

Ejemplo: Sea la matriz,

$$A = \begin{pmatrix} 7 & -6 \\ -8 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \end{pmatrix}$$

La solución exacta es $(x_1, x_2) = (1/5, -4/15)$. Usando la descomposición de Jacobi observamos que, generamos la iteración,

$$\begin{pmatrix} 7 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{pmatrix} = \begin{pmatrix} 0 & 6 \\ 8 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \end{pmatrix} + \begin{pmatrix} 3 \\ -4 \end{pmatrix}$$

Entonces,

$$7x_1^{(k+1)} = 6x_2^{(k)} + 3$$
$$9x_2^{(k+1)} = 8x_1^{(k)} + 3$$

Tendriamos que intentar resolver dos iteraciones en caso de que la solución exista y sea exactamente la del sistema lineal.

Gauss-Seidel: El siguiente método es el de Gauss-Seidel, este consiste en separar de igual forma la matriz como antes $A = -L + D - U$, sin embargo, estabemos hacemos la siguientes descomposición con D invertible,

$$x = (D - L)^{-1}Ux + (D - L)^{-1}b$$

Formando la iteración,

$$(D - L)x^{(k+1)} = Ux^{(k)} + b$$

para todo $k = 0, 1, 2, \ldots$

Nota: Si D es invertible, entonces $D - L$ también lo es. Esto fácilmente se verifica puesto que $\det(D) = \det(D - L)$.

Lema (Serie de Neumann): *Sea $T \in \mathbb{R}^{N\times N}$ con $\rho(T) < 1$. Entonces $(I - T)$ es regular y,*

$$(I - T)^{-1} = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j$$

Teorema: *La sucesión $\{x^k\}_{k\in\mathbb{N}}$ definida por,*

$$x^{(k+1)} = Tx^{(k)} + c$$

converge para cada vector inicial x^0 a la única solución $x = Tx + c$ si y sólo si $\rho(T) < 1$.

Demostración: Supongamos que $\rho(T) < 1$. Sea x al única solución del sistema $x = Tx + c$ (bien definido puesto que $I - T$ es invertible). Luego al resta $x = Tx + c$ a la iteración, se obtiene que,

$$x^{(k+1)} - x = T(x^{(k)} - x)$$

Tomando $y^{(k+1)} := x^{(k+1)} - x$ se reduce a,

$$y^{(k+1)} = Ty^{(k)}$$

En particular,

$$y^{(k+1)} = T^{k+1}y^{(0)}$$

Entonces para cualquier $y^{(0)}$ se cumple que $\|T^{k+1}\| \to 0$ para $k \to \infty$. Por tanto $x^k \to x$ cuando $k \to \infty$.

Supongamos ahora que $x^{(k)}$ converge a x que es única solución de $x = Tx + c$ para cualquier valor inicial $x^{(0)}$. Esto implica que $I - T$ es invertible y luego T no tiene valor propio 1. Haciendo el mismo método de restar $x = Tx + c$ y definir $y^{(k)}$, observamos que,

$$\|T^{k+1}y^{(0)}\| = \|y^{(k+1)}\| \to 0$$

cuando $k \to \infty$, esto implica que $T^{k+1}y^{(0)} \to 0$ cuando $k \to \infty$ para todo $y^{(0)} \in \mathbb{R}^N$, es s decir, necesariamente $\rho(T) < 1$.

■

Corolario: *Sea $T \in \mathbb{R}^{N\times N}$ con $\|T\| < 1$ para una norma matricial inducida $\|\cdot\|$. Entonces $x^{(k+1)} = Tx^{(k)} + c$ converge para cada $x^{(0)} \in \mathbb{R}^N$ a la solución $x = Tx + c$, entonces,*

$$\|x - x^{(k)}\| \leq \|T\|^k \|x - x^{(0)}\|$$

y

$$\|x - x^{(k)}\| \leq \frac{\|T\|}{1 - \|T\|}\|x^{(k)} - x^{(k-1)}\|$$

Demostración: Si $\|T\| < 1$, entonces se cumple que $\|T^k\| \to 0$ cuando $k \to \infty$. Es decir, $\rho(T) < 1$ y por tanto $x^{(k)}$ converge a x para cualquier valor inicial $x^{(0)}$. Ahora, esto implica que,

$$\begin{aligned}\|x - x^{(k)}\| &= \|Tx + c - Tx^{(k-1)} - c\| \\ &= \|T(x - x^{(k-1)})\| \leq \|T\|\|x - x^{(k-1)}\|\end{aligned}$$

Haciendo este proceso $k - 1$ veces más, se obtiene que,

$$\|x - x^{(k)}\| \leq \|T\|^k \|x - x^{(0)}\|$$

Para al otra desigualdad notemos que $x^{(k+2)} - x^{(k+1)} = T(x^{(k+1)} - x^{(k)})$, entonces,

$$\begin{aligned}\|T\|\|x^{(k+1)} - x^{(k)}\| &\geq \big|\|T\|\|x - x^{(k)}\| - \|T\|\|x - x^{(k+1}\|\big| \\ &\geq \|x - x^{(k+1)}\| - \|T\|\|x - x^{(k+1)}\| = (1 - \|T\|)\|x - x^{(k+1)}\|\end{aligned}$$

De aquí se concluye la desigualdad. ■

Ejemplo: Consideremos el sistema lineal,

$$\begin{pmatrix} 7 & -6 \\ -8 & 9 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ -4 \end{pmatrix}$$

Por Jacobi tenemos que,

$$T = D^{-1}(L + U) = \begin{pmatrix} 0 & 6/7 \\ 8/9 & 0 \end{pmatrix}$$

Observemos que $\rho(T) < 1$, entonces la iteración converge a $x = Tx + c$. Y se cumple que,

$$\|x - x^{(k)}\|_\infty \leq \left(\frac{8}{9}\right)^8 \|x - x^{(0)}\|_\infty$$

donde $\|T\|_\infty = 8/9$. Por tanto mediante Jacobi la iteración converge.

Definición: *Una matriz $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ se llama **estrictamente diagonal dominante** si,*

$$|a_{ii}| > \sum_{j=1, j\neq i}^{N} |a_{ij}|$$

para todo $i = 1, \ldots, N$.

Observación: Otra forma de decir que una matriz es estrictamente diagonal dominante si el elemento positivo de la diagonal es mayor estricamente a la suma de los elementos de la fila sin contar el de la diagonal.

Teorema: *Los métodos de Jacobi y de Gauss-Seidel convergen para matrices estrictamente diagonal dominantes.*

Demostración: Sea A una matriz estricamente diagonal dominante. Se tiene la iteración,

$$x^{(k+1)} = Tx^{(k)} + c$$

- **Jacobi:** Observemos que D es invertible puesto que necesariamente $|a_{ii}| > 0$ para todo $i = 1, \ldots, N$. Luego $T = D^{-1}(L + U)$.Supongamos que para algún i, la suma $\sum_{j=1}^{N} |a_{ij}|$ sin incluir $j = i$, es no nulo, entonces se tiene que,

  $$\|D^{-1}(-L - U)\|_\infty = \max_{1\leq i\leq N} \sum_{j=1, j\neq i}^{N} \frac{|a_{ij}|}{|a_{ii}|} < 1$$

  Por tanto se tiene convergencia mediante Jacobi.

- **Gauss-Seidel por ver falta analizar el caso de A diagonal.**

Teorema (Stein-Rosenberg): *Sea $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ con $a_{ii} > 0$ para todo $i = 1, \ldots, n$. Y $a_{ij} \leq 0$ para todo ineqj. Sean T_J y T_{GS} las matrices de iteración correspondiente de los métodos de Jacobi y Gauss-Seidel respectivamente. Se verifica una de las siguientes relaciones:*

(i) $0 \leq \rho(T_{GS}) < \rho(T_J) < 1$.

(ii) $1 < \rho(T_J) < \rho(T_{GS})$.

(iii) $\rho(T_{GS}) = \rho(T_J) = 0$.

(iv) $\rho(T_{GS}) = \rho(T_J) = 1$.

Definición: *Sea $A \in \mathbb{C}^{N\times N}$ una matriz compleja.*

(i) Se define $A^* := \overline{A}^T$*, o sea,*

$$a^*_{ij} := \overline{a}_{ji}$$

(ii) La matriz A *se llama* ***Hermitiana*** *si,*

$$A^* = A$$

Observación:

- Si $A \in \mathbb{R}^{N\times N}$ entonces A es hermitiana si y sólo si es simétrica.
- Si A es una matriz hermitiana. entonces para todo $z \in \mathbb{C}^N$ se tiene que $z^*Az \in \mathbb{R}$.

Teorema: *Si* $A \in \mathbb{C}^{N\times N}$ *es hermitiana y definida positiva* ($z^*Az > 0$ *para todo* $z \neq 0$*). Entonces el método de Gauss-Seidel es convergente.*

Ejemplo: Sea la matriz,

$$A = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}$$

Este es definida positiva puesto que,

$$\det(A_1) = 1, \quad \det(A_2) = \frac{3}{4}, \quad \det(A_3) = \frac{1}{2}$$

Además es simétrica. Entonces A es una matriz hermitiana definida positiva y por tanto converge con Gauss-Seidel.

Veamos que pasa con el método de Jacobi. Observemos que la transformación de la iteración es,

$$T_J = -\begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

Luego,

$$\det(\lambda I - T_J) = (\lambda + 1)\left(\lambda - \frac{1}{2}\right)^2$$

Entonces $\rho(T_J) = 1$, por tanto Jacobi no converge.

Observación: La diagonal de A es dominante pero no estricamente dominante.

Ejemplo:

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

terminar

3. Valores Propios

Oscilador Armónico: Sabemos que se cumple la siguiente ley de Hooke:

$$F = -ky$$

Donde F es la fuerza, k la constante elástico y y el desplazamiento desde la elongación natural. Por la segunda ley de Newton $F = ma = md^2y/dt^2$ se cumple que,

$$m\frac{d^2y}{dt^2} = -ky$$

con solución $y = Ae^{i\omega t}$ y $\omega = \sqrt{k/m}$ (frecuencia angular).

En tres dimensiones se cumple que,

$$-\Delta u(x) := -\left(\frac{\delta^2}{\delta x_1^2} + \frac{\delta^2}{\delta x_2^2} + \frac{\delta^2}{\delta x_3^2}\right) u(x) = \lambda u(x)$$

donde $x \in \Omega \subseteq \mathbb{R}^3$ para algún valor λ. Nuestro objetivo es estudiar λ.

La ecuación $-\Delta u = \lambda u$ (usando que div∇u) se reescribe como,

$$\int_\Omega \nabla u \cdot \delta v dx = \lambda \int_\Omega uv dx$$

con $u \in Vs \setminus \{0\}$ para todo $v \in V$ y $\lambda \in \mathbb{R}$ una constante real.

Podemos aproximar u por $u_N \in V_N \subset V$ donde $\dim(V_N) = N < \infty$ con base $\langle\{\phi_1, \ldots, \phi_N\}\rangle$, o sea $u_N = c_1\phi_1 + \cdots + c_N\phi_N$ se determina por,

$$\int_\Omega \nabla u_N \nabla \phi_j j dx = \lambda_N \int_\Omega u_N \phi_j dx, \quad j = 1, \ldots, N$$

$$\Longleftrightarrow$$

$$\sum_{i=1}^N c_i \int_\Omega \nabla\phi_i \nabla\phi_j dx = \lambda_N \sum_{i=1}^N c_i \int_\Omega \phi_i\phi_j dx, \quad j = 1, \ldots, N$$

$$\Longleftrightarrow$$

$$A_1\vec{c} = \lambda_N A_2\vec{c}$$

donde,

$$A_1 = \left(\int_\Omega \nabla\phi_j\nabla\phi_i dx\right)_{ij}, \quad A_2 = \left(\int_\Omega \phi_j\phi_i dx\right)_{ij}, \quad \vec{c} = (c_1, \ldots, c_N)^T$$

Obteniendo un problema de valores propios.

Definición: *Sea $A \in \mathbb{C}^{N\times N}$ una transformación. Decimos que $\lambda \in \mathbb{C}$ es un **valor** propio si existe un vector $x \in \mathbb{C}^N$ no nulo tal que $Ax = \lambda x$. Al vector x le decimos **vector propio**.*

Observaciones:

- $p(\lambda) := \det(A - \lambda I)$ es el polinomio característico de A. Tiene grado N y los valores propios de A son sus raíces.
- El cálculo de $\det(A - \lambda I) = 0$ es caro pero no imposbile, recordemos que no hay fórmulas explícitas para raíces de polinomios de grado mayor o igual a 5.
- Sean $\lambda_1, \ldots, \lambda_k$ son los valores propios de A con vectores propios x_k donde no se repiten los valores propios, se tiene que $x_1, \ldots, x_k$ son linealmente independientes.

Definición: *Sea* $\{x^{(1)}, \ldots, x^{(k)}\}$ *una colección de vectores. Decimos que* ***ortogonal(ortonormal)*** *si,*

$$(x^{(i)})^T x^{(j)} = c_{ij}\delta_{ij}$$

(es ortonormal si $c_{ij} = 1$*). Donde* δ_{ij} *es la delta de Kronecker.*

Definición: *Sea* $Q \in \mathbb{R}^{N\times N}$ *una matriz. Decimos que es ortogonal si* $Q^{-1} = Q^T$*.*

Ejemplo: Las matrices de permutación son ortogonales.

Observación: Las matrices simétricas reales son diagonizables por matrices ortogonales. En particular, si $A = A^T$ matriz real, entonces existe Q matriz ortogonal tal que $A = Q^T DQ$ donde D es una matriz diagonal con elementos los valores propios de A.

Teorema (Gershgorin): *Sea* $A \in \mathbb{R}^{N\times N}$ *una matriz real. Entonces los valores propios de* A *están en,*

$$\bigcup_{i=1}^{N} R_i \quad \textit{donde} \quad R_i = \{z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{N} |a_{ij}|$$

Si k *de los círculos* R_i*, no tienen intersección con los deas* $(n - k)$ *círculos, entonces los* k *círculos contienen* k *valores propios (contando según multiplicidad).*

Ejemplo: Consideremos la matriz,

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{pmatrix}$$

El cual es una matriz real, entonces,

$$R_1 = \{z \in \mathbb{C} : |z - 4| \leq 2\}$$
$$R_2 = \{z \in \mathbb{C} : |z - 2| \leq 1\}$$
$$R_3 = \{z \in \mathbb{C} : |z - 9| \leq 2\}$$

terminar

Observación: El teorema anterior se verficia para los círculos R_i' definidos de la siguiente forma:

$$R_i' = \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{N} |a_{ji}| \right\}$$

Es decir,

$$\bigcup_{i=1}^{N} R_i'$$

contiene todos los valores propios de A. Para ver esto observemos que A^T tiene los mismos valores propios de A, de forma que los círculos de A^T sobre las filas, es igual a los círculos de A sobre las columnas (R_i').

3.1. Método de Potencia

Sea $A \in \mathbb{R}^{N\times N}$ matriz real con N valores propios $\lambda_1, \ldots, \lambda_N$ con valores propios $v^{(1)}, \ldots, v^{(N)}$ linealmente independientes tales que $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N| \geq 0$. Sea $x \in \mathbb{R}^N$, entonces existen β_j constantes tales que,

$$x = \sum_{j=1}^{N} \beta_j v^{(j)}$$

Aplicando la transformación A obtenemos que,

$$Ax = \sum_{j=1}^{N} \beta_j \lambda_j v^{(j)}$$

Esto se puede hacer una infinidad de veces. Para la k-ésima iteración obtenemos que,

$$A^k x = \sum_{j=1}^{N} \beta_j (\lambda_j)^k v^{(j)}$$

Supongamos que λ_1 no es nulo, entonces,

$$A^k x = \lambda_1^k \sum_{j=1}^{N} \beta_j v^{(j)}$$

Si β_1 no es nulo, entonces al tomar k suficientemente grande, se tiene que $A^k x \sim \lambda_1^k \beta_1 v^{(1)}$ debido a que ordenamos los valores propios.

Sea $x^{(0)} \in \mathbb{R}^N$ con $x_{p_0}^{(0)} = \|x^{(0)}\|_\infty = 1$ y sea $y^{(1)} = Ax^{(0)}$. Entonces,

$$\begin{aligned}\mu^{(1)} := y_{p_0}^{(1)} = \frac{y_{p_0}^{(1)}}{x_{p_0}^{(1)}} &= \frac{\beta_1\lambda_1 v_{p_0}^{(1)} + \sum_{j=2}^{N} \beta_j \lambda_j v_{p_0}^{(j)}}{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^{N} \beta_j v_{p_0}^{(j)}} \\ &= \lambda_1 \frac{\lambda_1 v_{p_0}^{(1)} + \sum_{j=2}^{N} \beta_j \frac{\lambda_j}{\lambda_1} v_{p_0}^{(j)}}{\beta_1 v_{p_0}^{(1)} + \sum_{j=2}^{N} \beta_j v_{p_0}^{(j)}}\end{aligned}$$

Sea p_1 el menor entero tal que $|y_{p_1}^{(1)}| = \|y^{(1)}\|_\infty$, luego,

$$x^{(1)} := \frac{y^{(1)}}{y_{p_1}^{(1)}} = \frac{Ax^{(0)}}{y_{p_1}^{(1)}}$$

Entonces $x_{p_1}^{(1)} = \|x^{(1)}\|_\infty = 1$ e $y^{(2)} = Ax^{(1)} = \frac{A^2x^{(0)}}{y_{p_1}^{(1)}}$. Luego replicamos el proceso con $\mu^{(2)}$ y así sucesivamente.

Finalmente obtenemos cuatro sucesiones $x^{(n)}, y^{(n)}, p_n, \mu^{(n)}$ donde,

$$\begin{aligned}
y^{(n)} &:= Ax^{(n-1)} \\
\mu^{(n)} &:= y_{p_{n-1}}^{(n)} = \lambda_1 \frac{\lambda_1 v_{p_{n-1}}^{(1)} + \sum_{j=2}^{N} \beta_j \frac{\lambda_j}{\lambda_1} v_{p_{n-1}}^{(j)}}{\beta_1 v_{p_{n-1}}^{(1)} + \sum_{j=2}^{N} \beta_j v_{p_{n-1}}^{(j)}} \\
p_n &: |y_{p_n}^{(n)}| = \|y^{(n)}\|_\infty \\
x^{(n)} &:= \frac{y^{(n)}}{y_{p_n}^{(n)}} = \frac{A^n x^{(0)}}{\prod_{k=1}^{n} y_{p_k}^{(k)}}
\end{aligned}$$

Si β_1 no es nulo, entonces $\mu^{(n)} \to \lambda_1$ y $x^{(n)} \to x$ para $n \to \infty$ donde x tiene valor propio λ y $\|x\|_\infty = 1$. Este método se llama **Método de potencia** o de **von Mises**.

Ejemplo: Consideremos la siguiente matriz,

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{pmatrix}$$

Determinemos el valor propio dominante usando el método de potencias.

Observación: Se presentan las siguientes dificultades:

- A veces no se sabe si la matriz tiene valor propio dominantes simple.
- Encontrar el vector inicial tal que $\beta_1 \neq 0$ (esto puede producir pertubaciones).

Convergencia: Consideremos las hipótesis del método de potencia. Se tiene que,

$$|\mu^{(n)} - \lambda_1| = \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)$$

para $n \to \infty$, y,

$$\frac{|\mu^{(n+1)} - \lambda_1|}{|\mu^{(n)} - \lambda_1|} \approx \left|\frac{\lambda_2}{\lambda_1}\right| < 1$$

Siendo una convergencia lineal.

3.2. Cociente de Rayleigh y Método de la Potencia

Definición: *Sea $A \in \mathbb{R}^{N\times N}$. Se define el* ***coeficiente de Rayleigh de*** *A , la expresión,*

$$R_A(x) := \frac{x^*Ax}{x^*x}$$

para todo $x \in \mathbb{C}^N \setminus \{0\}$ y donde $x^* = \overline{x}^T$.

Observación: Si A es una matriz normal ($A^TA = AA^T$), entonces el conjunto $\{z \in \mathbb{C} : R_A(x) = z, \text{para algún } x \in C^N \text{ no nulo}\}$ es una envoltura convexa en $\mathbb{C}$ de los valores propios de A.

En conclusión, sea x_λ vector propio del valor propio λ, se tiene que $R_A(x_\lambda) = \lambda$. Si además A es simétrica, entonces,

$$\lambda_{\text{mín}}(A) = \min_{x\in\mathbb{R}^N\setminus\{0\}} \frac{x^TAx}{x^Tx}$$
$$\lambda_{\text{máx}}(A) = \max_{x\in\mathbb{R}^N\setminus\{0\}} \frac{x^TAx}{x^Tx}$$

Podemos volver aplciar el método de potencia pero con respecto al cociente de Rayleigh. En este caso se generan cinco sucesiones $x^{(n)}, y^{(n)}, p_n, \mu^{(n)}, \hat{\mu}^{(n)}$. Donde,

$$\begin{aligned} y^{(n)} &:= Ax^{(n-1)} \\ \mu^{(n)} &:= \frac{y^{(n)}}{x^{(n-1)}_{p_n-1}} \\ p_n &: \quad |y^{(n)}_{p_n}| = \|y^{(n)}\|_\infty \\ x^{(n)} &:= \frac{y^{(n)}}{\|y^{(n)}\|_2} \\ \hat{\mu}^{(n)} &:= R_A(x^{(n-1)}) = x^{(n-1)^T}y^{(n)} \end{aligned}$$

Si $A = A^T$ con vectores propios ortonormales $v^{(1)}, \ldots, v^{(N)}$ **terminar calulo**.

$$x^{(n-1)^T}y^{(n)} = \lambda_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2n-2}\right)$$

terminar

Ejemplo: Consideremos la matriz,

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 2 & 1 \\ -2 & 0 & 9 \end{pmatrix}$$

Determinemos el valor propio dominante pero usando el cociente de Rayleigh.

terminar

Ejemplo: Consideremos la matriz,

$$A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 3 & -2 \\ 1 & -2 & 3 \end{pmatrix}$$

Calculemos el valor propio dominante con el cociente de Rayleigh.

Teorema: *Sea $A \in \mathbb{R}^{N\times N}$ simétrica con valores propios $\lambda_1, \ldots, \lambda_N$. Si,*

$$\|Ax - \lambda x\|_2 < \varepsilon$$

para algún $x \in \mathbb{R}^N$ con $\|x\|_2 = 1$ y $\lambda \in \mathbb{R}$, entonces,

$$\min_{1\leq j\leq N} |\lambda_j - \lambda| < \varepsilon$$

3.3. Método de Potencia Inversa de Wielandt

Sea $A \in \mathbb{R}^{N\times N}$ una matriz real con valores propios $\lambda_1, \ldots, \lambda_N$ y $v^{(1)}, \ldots, v^{(N)}$ vectores propios linealmente independientes. Sea $q \in \mathbb{R}$ (o $\mathbb{C}$) tal que $q \neq \lambda_j$ con $j = 1, \ldots, N$. Entonces la matriz $(A - qI)^{-1}$ tiene valores propios:

$$\frac{1}{\lambda_1 - q}, \frac{1}{\lambda_2 - q}, \ldots, \frac{1}{\lambda_N - q}$$

¿Cómo lo determinamos? Usaremos el método de potencia aplicado a la matriz $(A - qI)^{-1}$. La sucesión $\mu^{(n)}$ con respecto a la matriz anterior converge a,

$$\mu := \frac{1}{\lambda_k - q}$$

donde k es maximal de los $1/|\lambda_j - q|$. Esto se reescribe de forma que,

$$q + \frac{1}{\mu^{(n)}} \to \lambda_k$$

para $n \to \infty$.

3.4. Transformaciones de Semejanza

Queremos transformar una matriz simétrica A a una matriz tridiagonal B. El método sigue en definir $A_0 := A$, luego mediante transformaciones invertibles, definir $A_1 := T_1^{-1} 1 A_0 T_1$. Eso hasta obtener una tridiagonal $B = A_m$ donde $T^{-1}AT$ con $T = T_1 T_2 \ldots T_m$.

En caso de que se pueda realizar estar, se tendría las siguientes propiedades (deseadas):

- B tiene los valores propios A.

- El cálculo de valores propios de B es más fácil que A.
- La transformación está bien condicionada:

$$B + \Delta B = T^{-1}(A + \Delta A)T$$

con $\Delta A = T\Delta T^{-1}$, luego $\|B\| \leq \|T^{-1}\|A\|\|T\| = \text{cond}(T)\|A\|$ y $\|\Delta A\| \leq \text{cond}(T)\|\Delta B\|$. Entonces,

$$\frac{\|\Delta A\|}{\|A\|} \leq \text{cond}(T)^2 \frac{\|\Delta B\|}{\|B\|}$$

Proyecciones: Empezemos por una observación.

Observación: Sea Q ortogonal. Entonces,

$$\|Q\|_2 = \max_{\|x\|_2=1} \|Qx\|_2 = \max_{\|x\|_2=1} \|x\|_2 = 1$$

Por tanto $\text{cond}_2(Q) = 1$

Ahora, sea $\omega \in \mathbb{R}^N$, se tiene que $\omega\omega^T x$ es la proyección de x en dirección ω. Si $\|\omega\|_2 = 1$ entonces $(I - \omega\omega^T)x$ es la proyección de x en dirección ortogonal a ω.

Definición: *Sea $\omega \in \mathbb{R}^N$ tal que $\|\omega\|_2 = 1$. Definimos la transformación de **Householder** por:*

$$P_\omega = I - 2\omega\omega^T \in \mathbb{R}^{N\times N}$$

Observación: $P_\omega x$ es la reflexión de x en el hiperplano $\{x \in \mathbb{R}^N : \omega^T x = 0\}$.

Ejemplo: Si $\omega = e_N$, entonces,

$$\{x \in \mathbb{R}^N : \omega^T x = 0\} = \{x \in \mathbb{R}^N : x_N = 0\}$$

Y,

$$P_\omega = \text{diag}(1, 1, \ldots, 1, -1)$$

Por lo tanto,

$$P_\omega x = \begin{pmatrix} x_1 \\ \vdots \\ x_{N-1} \\ -x_N \end{pmatrix}$$

es la reflexión de x en el hiperplano $\{x \in \mathbb{R}^N : x_N = 0\}$.

Teorema: *La transformación de Householder es simétrica y ortogonal.*

Demostración: Sea $\omega \in \mathbb{R}^N$ tal que $\|\omega\|_2 = 1$.

- **Simétrica:** Aplicando la transpuesta se observa que,

$$(P_\omega)^T = I - 2\omega\omega^T$$

  Claramente es simétrica.

- **Ortogonal:** Observemos que,

$$P_\omega P_\omega^T = (I - 2\omega\omega^T)(I - 2\omega\omega^T) = I - 4\omega\omega^T + 4\omega\omega^T = I$$

  Por tanto es ortogonal.

■

Ahora, volvamos a ver como transformar una matriz simétrica a una matriz tridiagonal usando transformaciones de Householder.

Primero queremos P_1 tal que,

$$P_1 A P_1 = \begin{pmatrix} \hat{a}_{11} & & \\ \hat{a}_{21} & \hat{a}_{22} & \\ 0 & & \\ \vdots & & \end{pmatrix}$$

Elegimos,

$$P_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \hat{P} & \\ 0 & & & \end{pmatrix}$$

donde $\hat{P} \in \mathbb{R}^{(N-1)\times(N-1)}$ donde—

Resume: Se tiene lo siguiente:

- $P_1 = I - \beta u u^T \in \mathbb{R}^{N\times N}$.
- $\beta = (\|x\|_2(|x_2| + \|x\|_2))^{-1}$.
- $u = (0, \operatorname{sgn}(x_2)(|x_2| + \|x\|_2), x_3, \dots, x_N)^T$.

Luego,

$$P_1 A P_1 = \begin{pmatrix} a_{11} & \alpha & 0 & \dots \\ \alpha & & & \\ 0 & & \hat{P}\hat{A}\hat{P} & \\ \vdots & & & \end{pmatrix}$$

Ahora repetimos el proceso para $\hat{P}\hat{A}\hat{P}$ y así sucesivamente obteneniendo $P_{N-2} \dots P_1 A P_1 \dots P_{N-2}$.

Conclusiones: Para $A = A^T$, la matriz $H := P_{N-2} \dots P_1 A P_1 \dots P_{N_2}$ es tridiagonal simétrica. Para A no simétrica, H es una matriz **Hessenberg** superior:

$$H = (h_{ij}), \quad h_{ij} = 0 \, (i \geq j + 2)$$

Hemos mostrado el siguiente resultado.

Teorema: *Cada matriz $A \in \mathbb{R}^{N \times N}$ tiene una descomposición QR:*

$$A = QR$$

Q ortognal y R triangular superior.

3.5. Método QR

Sea $A \in \mathbb{R}^{N \times N}$ matriz simétrica tridiagonal,

$$A = \begin{pmatrix} a_1 & b_2 & & 0 & \\ b_2 & a_2 & b_3 & & \\ & b_3 & a_3 & \ddots & \\ & & \ddots & \ddots & b_N \\ & 0 & & b_N & a_N \end{pmatrix}$$

Se presentan lo siguientes casos particulares con A reducible.

- Si $b_2 = 0$, entonces a_1 es valor propio.
- Si $b_N = 0$, entonces a_N es valor propio.
- Si $b_j = 0$ con $j = 3, \dots, N-1$. Entonces los valores propios de A son los de la matriz,

$$\begin{pmatrix} a_1 & b_2 & & 0 \\ b_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_{j-1} \\ 0 & & b_{j-1} & a_{j-1} \end{pmatrix} \quad \text{y} \quad \begin{pmatrix} a_j & b_{j+1} & & 0 \\ b_{j+1} & \ddots & \ddots & \\ & \ddots & \ddots & b_N \\ 0 & & b_N & a_N \end{pmatrix}$$

Consideremos $A \in \mathbb{R}^{N \times N}$ matriz real simétrico tridiagonal irreducible, es decir,

$$A = \begin{pmatrix} a_1 & b_2 & & 0 & \\ b_2 & a_2 & b_3 & & \\ & b_3 & a_3 & \ddots & \\ & & \ddots & \ddots & b_N \\ & 0 & & b_N & a_N \end{pmatrix}, \quad b_j \neq 0, \; j = 2, \dots, N$$

Método QR: Sea $A^{(1)} := A$. Para $i \geq 1$ se tiene $A(i) = Q^{(i)} R^{(i)}$ (descomposición QR) y $A^{(i+1)} := R^{(i)} Q^{(i)}$ (transformación QR).

Propiedades:

- $A^{(i+1)}$ es semejante a $A^{(i)}$.
- $A^{(i+1)}$ es tridiagonal si $A^{(i)}$ lo es.
- La transformación QR es estable.

Nota: El método QR se aplica también a matrices de tipo Hessenberg superior.

Teorema: *Sea $A \in \mathbb{R}^{N\times N}$ diagonalizable con valores propios separados:*

$$|\lambda_1| > |\lambda_2| > \dots |\lambda_N|$$

Además, si la matriz X de la diagonalización $A = XDX^{-1}$ con $D = diag(\lambda_1, \dots, \lambda_N)$ tiene descomposición LU. Entonces la sucesión $A^{(k)} = (a_{ij}^{(k)})$ generada por el método QR satisface,

$$\lim_{k\to\infty} a_{ii}^{(k)} = \lambda_i$$

con $i = 1, \dots, N$. Con razones de convergencia,

$$\mathcal{O}\left(\left|\frac{\lambda_{k+1}}{\lambda_i}\right|^k\right),\ i < N \qquad \mathcal{O}\left(\left|\frac{\lambda_N}{\lambda_{N-1}}\right|^k\right),\ i = N$$

Si X no tiene descomposición LU, entones los elementos diagonales $a_{ii}^{(k)}$ todavía convergen a los valores propios de A, pero no necesariamente en el orden anterior.

Ejemplo:

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 6 & 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 8 & 3 & 0 & 0 & 0 \\ 0 & 0 & 3 & 10 & 4 & 0 & 0 \\ 0 & 0 & 0 & 4 & 12 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 & 14 & 6 \\ 0 & 0 & 0 & 0 & 0 & 6 & 16 \end{pmatrix}$$

4. Ecuaciones no Lineales

4.1. Método de Bisección

Problema: Consideremos una función $f : \mathbb{R} \to \mathbb{R}$ (o $f : [a, b] \to \mathbb{R}$). Queremos encontrar $r \in \mathbb{R}/[a, b]$ tal que $f(r) = 0$.

Una posible solución es usar el siguiente teorema.

Teorema (Bolzano, Punto medio, Valor Intermedio): *Sea $f : [a, b] \to \mathbb{R}$ función continua con $f(a)f(b) < 0$. Entonces f tiene una raíz en (a, b).*

Sea $f : [a, b] \to \mathbb{R}$ continua tal que $f(a)f(b) < 0$. Podemos usar el teorema anterior, sin embargo, esto nos entrega la existencia de la raíz, y nos interesa determinar la raíz. Calculemos $c = (a + b)/2$. Si $f(a) > 0$ y $f(b) < 0$, entonces se cumple los siguientes tres casos:

- $f(c) = 0$. En este caso c es raíz y terminamos el estudio.
- $f(c) < 0$, entonces debería haber una raíz en el intervalo (a, c).
- $f(c) > 0$, entonces debería haber una raíz en (c, b).

Podemos continuar con el intervalo que contiene una raíz y así sucesivamente. De esta forma, se genera una sucesión,

$$\begin{aligned} x_1 &:= \frac{a+b}{2} \in [a_1, b_1] \\ x_2 &\in [a_2, b_2] \\ &\vdots \\ x_n &\in [a_n, b_n] \end{aligned}$$

Donde,

$$\begin{aligned} b_1 - a_1 &= \frac{b-a}{2} \\ b_2 - a_2 &= \frac{b_1 - a_1}{2} = \frac{b-a}{4} \\ &\vdots \\ b_n - a_n &= \frac{b-a}{2^n} \end{aligned}$$

y se satisface que,

$$\begin{aligned} a \leq a_1 \leq a_2 \leq \cdots \leq b \\ b \geq b_1 \geq b_2 \geq \cdots \geq a \end{aligned}$$

Esto implica que existen los límites,

$$r_1 := \lim_{n\to\infty} a_n, \quad r_2 = \lim_{n\to\infty} b_n$$

que son iguales $L = r_1 = r_2$. Por continuidad se tiene que,

$$f(a_n)f(b_n) \leq 0 \overset{n\to\infty}{\longrightarrow} f(L)^2 \leq 0$$

Por lo tanto $f(L) = 0$, es decir, hemos construido una sucesión que converge a una raíz de f. El procedimiento se conoce por **método de bisección**. Se resume en el siguiente resultado.

Teorema: *Sea $f : [a, b] \to \mathbb{R}$ continua tal que $f(a)f(b) < 0$. Entonces el método se bisección genera una sucesión x_n tal que,*

- *$x_n \to r$ para $n \to \infty$ donde $f(r) = 0$.*
- *$|r - x_n| \leq \frac{b-a}{2^n}$*

Ventajas del método: Basta con trabajar con una función continua y con cota de error precisa.

Desventaja del método: La convergencia es lenta y mejorable.

Ejemplo: Consideremos $f(x) = x^2 - 2$ que es continua. De antemano sabemos que una raíz es $\sqrt{2}$. Pensemos en $f : [1, 3] \to \mathbb{R}$ ya que $f(1)f(3) < 0$, de forma que $\sqrt{2}$ es la única raíz. Supongamos que no sabemos la raíz y necesitamos determinar mediante aproximaciones. Sea $x_1 = 2$, entonces $f(x_1) = 2 > 0$, luego consideramos el intervalo $[1, 2]$. Luego escogemos $x_2 = 3/2$, entonces $f(x_2) = 1/4 > 0$, luego consideramos el intervalo $[1, 3/2]$. Haciendo el proceso de forma reiterativa, obtenemos que,

$$f(a_n), f(b_n) \to f(\sqrt{2}) = 0$$

En particular, $x_8 = 1{,}4140625\ldots$ (tiene 4 cifras exactas de $\sqrt{2}$), y la estimación del error con respecto a x_8 es,

$$|\sqrt{2} - x_8| \leq \frac{3-1}{2^8} = \frac{1}{2^7}$$
$$\frac{|\sqrt{2} - x_8|}{\sqrt{2}} \leq \frac{1}{128\sqrt{2}} \leq 0,005 \approx 5 \cdot 10^{-3}$$

Es decir, x_8 tiene aproximadamente tres cifras significativas. Esta estimación nos dice que x_8 tiene al menos 3 cifras exactas, y además, el error se reduce por un factor $1/10$ en cada 3 o 4 pasos.

4.2. Iteración de Punto Fijo

Supongamos que r es una raíz de una función f, entonces $f(r) = 0$. Sumando r a ambos lado obtenemos $f(r) + r = r$ y definiendo $g(x) := f(x) + x$, vemos que r es un punto fijo de la función g. En estos casos usamos el siguiente resultado.

Teorema (Punto fijo de Banach): *Sea $g : [a, b] \to [a, b]$ función continua. Entonces un único punto fijo $p \in [a, b]$ tal que $g(p) = p$. Si además g es una concentración, es decir, existe $L < 1$ tal que $|g(x) - g(y)| \leq L|x - y|$ para todo $x, y \in [a, b]$, entonces el punto fijo es único en $[a, b]$.*

Ejemplo: Sea la función $g(x) = (x^3-1)/3$, observemos que $g'(x) = 2x/3$. Tomando $x \in [-1,1]$ se verifica que $g(x) \in [-1,1]$ y que,

$$|g(x) - g(y)| = |g'(\xi)||x-y| \leq \frac{2}{3}|x-y|$$

para todo $x, y \in [-1,1]$. Por lo que g es una contracción y tiene único punto fijo en $[-1,1]$.

Iteración de Picard: Dado p_0 valor inicial cualquiera, definimos la sucesión $p_{n+1} := g(p_n)$ para $n = 0, 1, \ldots$

Teorema: *Sea $g : [a,b] \to [a,b]$ una concentración de constante L. Entonces para todo valor inicial p_0 de la iteración de Picard, se tiene que la iteración converge a un único punto fijo p en $[a,b]$. Además,*

$$|p - p_n| \leq L^n \,\text{máx}\{p_0 - a, b - p_0\}$$
$$|p - p_n| \leq \frac{L^n}{1-L}|p_1 - p|$$

Ejemplo: Sea $f(x) = x^3+4x-10$. Queremos una raíz en el intervalo $[1,2]$. Primero observemos que $f(1) = -5, f(2) = 14$, es decir $f(1)f(2) < 0$, por lo que hay al menos una raíz en $[1,2]$. Sea $g(x) := f(x) + x = x^3 + 4x^2 + x - 10$, observemos que $g'(x) = 3x^2 + 8x + 1$, en particular $g'(x) \geq 1$ para todo $x \in [1,2]$, entoces no podemos el teorema de punto fijo de Banach. Para esto podemos reescribir el problema, podemos definir las funcione:

$$g_1(x) := \left(\frac{10}{x} - 4x\right)^{1/2}$$
$$g_2(x) := \frac{1}{2}(10 - x^3)^{1/2}$$
$$g_3(x) := \left(\frac{10}{4+x}\right)^{1/2}$$
$$g_4(x) := x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x} = \frac{2x^3 + 4x^2 + 10}{3x^2 + 8x}$$

Que están bien definidos en $[1,2]$ y que si determinamos un punto fijo, tenemos raíz de $f(x)$. Observemos que,

$$g_1'(x) = \frac{1}{2}\left(\frac{10}{x} - 4x\right)^{-1/2}\left(\frac{-10}{x^2} - 4\right)$$
$$g_2'(x) = \frac{1}{4}(10 - x^3)^{-1/2}(-3x^2)$$
$$g_3'(x) = \frac{1}{2}\left(\frac{10}{4+x}\right)^{-1/2}\left(-\frac{10}{(4+x)^2}\right)$$
$$g_4'(x) = \frac{6x^4 + 40x^3 + 32x^2 - 60x - 80}{9x^4 + 48x^3 + 64x^2}$$

Observemos que,

$$1 \leq g_4(x) \leq 2, \quad g_4'(x) < 1$$

para todo $x \in [1, 2]$. **terminar...**

4.3. Método de Newton:

Vamos a desarrolar el método de Newton.

Método de Newton (Newton-Raphson): Sea $f : [a, b] \to \mathbb{R}$ derivable con raíz $r \in (a, b)$. Sea $x_0 \in [a, b]$ tal que $f'(x_0) \neq 0$. Sea x_1 tal que.

$$f(x_0) + (x_1 - x_0)f'(x_0) = 0 \iff x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

De forma reiterativa, dado x_n, podemos tomar x_{n+1} tal que,

$$f(x_n) + (x_{n+1} - x_n)f'(x_n) = 0 \iff x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

donde $f'(x_n) = 0$. Obteniendo una sucesión $\{x_n\}_{n \in \mathbb{N}_0}$. Análicemos la convergencia.

Suposiciones: Sea $f :\in C^2[a, b]$ y sea $r \in [a, b]$ raiz simple de f tal que $f'(r) \neq 0$. Sea $e_n := x_n - r$, entonces,

$$e_{n+1} = x_{n+1} - r = \frac{e_n f'(x_n) - f(x_n)}{f'(x_n)}$$

terminar.

Conclusiones:

- El método de Newton converge, localmente, de manera **cuadrática**.
- Por análisis como iteración de Picard: constante Lipshitz/de contracción L.

  2

  bajo las supociociones del teorema, o sea $L \to 0$ cerca de la raíz. Este efecto es la convergencia cuadrática.
- Valor inicial con bisección, por ejemplo, o estimaciones gráficas.
- Desventaja del método: calcular derivadas.

Ejemplo: Sea $f(x) = x^2 - 2$. Mediante el método de bisección debemos estudiar $f : [1, 3] \to \mathbb{R}$ donde $f(1)f(3) = -7 < 0$, aunque no es efectivo. Pero mediante el método de newton es diferente. Se tiene que,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{x_n}{2} + \frac{1}{x_n}$$

Ejemplo: Calculemosla raíz de $f(x) = x^3$, que es claramente diferenciable, entonces,

$$x_{n+1} = x_n + \frac{f(x_n)}{f'(x_n)} = \frac{2}{3}x_n$$

Esto implica que $x = 0$ es la raíz, en particular $|e_{n+1}| = \frac{2}{3}|e_n|$ es decir, es una convergencia lineal .

Ejemplo: Sea la función,

$$f(x)$$

Definición: *Sea $f : (a, b) \to \mathbb{R}$ una función. Decimos que es* ***convexa*** *si,*

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

para todo $x, y \in (a, b)$ y para todo $\alpha \in (0, 1)$.

Observación: Si $f \in C^2[a, b]$ tal que $f'' > 0$, entonces es convexa.

Teorema: *Sea $f \in C^2(\mathbb{R})$ tal que $f'' > 0$ y sea x^* el punto mínimo de f. Entonces, para cada valor inicial $x_0 \neq x^*$, el método de Newton converge a una raíz de f si existe.*

Observación: Siguiendo la demostración $f \in C^2(\mathbb{R})$ y $f'' > 0$ puede reducirse a $f \in C^2[a, b]$ y $f'' > 0$ en $[a, b]$, mientras las iteraciones se mantengan en el intervalo $[a, b]$.

4.4. Sistemas de Ecuaciones no Lineles

Consideremos un sistema de n ecuaciones no lineales:

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ f_2(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n) &= 0 \end{aligned}$$

Y definimos,

$$F(x_1, \ldots, x_n) := \begin{pmatrix} f_1(\ldots) \\ f_2(\ldots) \\ \vdots \\ f_n(\ldots) \end{pmatrix} = 0 \in \mathbb{R}^n$$

que es una función $F : \mathbb{R}^n \to \mathbb{R}^n$ con funciones coordenadas $f_i : \mathbb{R}^n \to \mathbb{R}$ para $i = 1, \ldots, n$.

Ejemplo: Sea,

$$F(x_1, x_2) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} x_1^2 + x_2^2 - 4 \\ (x_1 - 2)^2 + x_2^2 - 4 \end{pmatrix}$$

Es un ejemplo de un sistema de 2 ecuaciones.

Teorema (Punto fijo de Banach): *Sea* $D = \{(x_1, \ldots, x_n)^T \in \mathbb{R}^n : a_i \leq x_i \leq b_i; i = 1, \ldots, n\} \subseteq \mathbb{R}^n\}$, *y sea* $G : D \to D$ *función continua. Entonces* G *tiene un punto fijo* $p \in D$. *Si además,* G *es contracción, es decir, existe* $L < 1$ *tal que* $\|G(x) - G(y)\| \leq L\|x - y\|$ *para todo* $x, y \in D$, *entonces el punto fijo es único y,*

$$x^{(k+1)} := G(x^{(k)})$$

para todo $k = 0, 1, \ldots$; *converge a* p *para todo valor inicial* $x^{(0)} \in D$ *y,*

$$\|x^{(k)} - p\| \leq \frac{L^k}{1 - L}\|x^{(1)} - x^{(0)}\|$$

donde $\|\cdot\|$ *es una normal vectorial.*

Lema: *Sea* $G = (g_1, \ldots, g_n)^T$. *Si las derivadas parciales son continuas tal que,*

$$\left|\frac{\partial g_i}{\partial x_j}(x)\right| \leq \frac{K}{n}$$

para todo $x \in D$ *con* $K < 1$, *entonces* $\|G(x) - G(y)\|_\infty \leq K\|x - y\|_\infty$ *para todo* $x, y \in D$.

Ejemplo: Consideremos el siguiente sistema de tres ecuaciones:

$$\begin{aligned} 3x_1 - \cos(x_2 x_3) - \frac{1}{2} &= 0 \\ x_1^2 - 81(x_2 + 0,1)^2 + \operatorname{sen}(x_3) + 1,06 &= 0 \\ e^{-x_1 x_2} + 20x_3 + \frac{10\pi - 3}{3} &= 0 \end{aligned}$$

Definimos g_1, g_2, g_3 de la siguiente forma:

$$\begin{aligned} g_1(x_1, x_2, x_3) &= \frac{1}{3}\cos(x_2 x_3) + \frac{1}{6} = x_1 \\ g_2(x_1, x_2, x_3) &= \frac{1}{9}\sqrt{x_1^2 + \operatorname{sen}(x_3) + 1,06} - 0,1 = x_2 \\ g_3(x_1, x_2, x_3) &= -\frac{1}{20}e^{-x_1 x_2} - \frac{10\pi - 3}{60} = x_3 \end{aligned}$$

Luego obtenemos la función,

$$\begin{pmatrix} g_1(x_1, x_2, x_3) \\ g_2(x_1, x_2, x_3) \\ g_3(x_1, x_2, x_3) \end{pmatrix} = \begin{pmatrix} \frac{1}{3}\cos(x_2 x_3) + \frac{1}{6} \\ \frac{1}{9}\sqrt{x_1^2 + \operatorname{sen}(x_3) + 1,06} - 0,1 \\ -\frac{1}{20}e^{-x_1 x_2} - \frac{10\pi - 3}{60} \end{pmatrix} : D := [-1, 1]^3 \to D$$

Se observa que $\left|\frac{\partial g_i}{\partial x_j}\right| \leq 0{,}281$ para todo i, j y para todo $x \in D$.

Fatla

Teorema: *Sea* $p \in \mathbb{R}^n$ *un punto fijo de* $G : \mathbb{R}^n \to \mathbb{R}^n$. *Supongamos que existe* $\delta > 0$ *tal que,*

(i) La derivada parcial $\partial g_i / \partial x_j$ *es continua en* N

4.5. Método de Newton

Recordemos la siguiente iteración escalar. Sea la iteración de Picard para:

$$g(x) = x - \phi(x)f(x)$$

donde $\phi(x) = (f'(x))^{-1}$. ¿Cómo definir la derivada de $F(x) = (f_1(x), \ldots, f_n(x))^T$ con $x \in \mathbb{R}^n$.

Sea,

$$A(x) = \begin{pmatrix} a_{11}(x) & \ldots & a_{1n}(x) \\ \vdots & \ldots & \vdots \\ a_{n1}(x) & \ldots & a_{nn}(x) \end{pmatrix}$$

donde $a_{ij} : \mathbb{R}^n \to \mathbb{R}$, y sea $G(x) = x - A(x)^{-1}F(x)$.

Teorema: *Sea $p \in \mathbb{R}^n$ un punto fijo de la función $G : \mathbb{R}^n \to \mathbb{R}^n$. Supongamos que existe $\delta > 0$ tal que,*

(i) $\frac{\partial g_i}{\partial x_j}$ *continua en* $N_\delta := \{x \in \mathbb{R}^n : \|x - p\|_\infty < \delta\}$ *para todo* i, j.

(ii) $\frac{\partial^2 g_i}{\partial x_j \partial x_k}$ *continua en* N_δ *y* $\left|\frac{\partial^2 g_i}{\partial x_j \partial x_k}\right| \leq M$ *en* N_δ *para todo* i, j, k.

(iii) $\frac{\partial g_i}{\partial x_j}(p) = 0$ *para todo* i, j.

Entonces existe $\hat{\delta} \leq \delta$ tal que,

$$x^{(k+1)} = G(x^{(k)}) \to p$$

para $k \to \infty$ de forma cuadráticamente para todo $x^{(0)} \in N_\delta$. Además,

$$\|x^{(k)} - p\|_\infty \leq \frac{n^2 M}{2}\|x^{(k-1)} - p\|_\infty^2$$

para todo $k \geq 1$.

Aplicando el teorema a

$$G(x) = x - A(x)^{-1}F(x)$$

resulta que,

$$A(x) = J_F(x) \text{ Matriz Jacobiana de } F$$

Método de Newton: Para un sistema de ecuaciones no lineales $F(x) = 0$ se tiene la iteración:

$$x^{(k+1)} = G(x^{(k)}), \quad G(x) := x - J_F^{-1}(x)F(x)$$

Bajo las condiciones del teorema anterior, esta iteración converge cuadráticamente a una raíz de F si $x^{(0)}$ está suficientemente cerca de la raíz.

Ayudantías

Ayudantía 1

P1:

Solución: Si una máquina tiene números másquina dado por $m = 5$ y $-16 \leq l \leq 16$, entonces significa que son los números reales tales que a lo más tienen 5 dígitos distintos de 0. Entonces la mejor aproximación es

$$0{,}27182 \cdot 10^1$$

El siguiente número de máquina es 0,27183. Por lo que la distancia de la aproximación del siguiente número de máquina es $0{,}00001 \cdot 10^1$.

P2:

Solución: Sea $f : U \to \mathbb{R}$ una función diferenciable, entonces una forma de estudiar la región donde puede estar mal condicionada, es estudiando el mapa,

$$K(x) := \left| \frac{xf'(x)}{f(x)} \right|$$

El cual es el número de condición relativo.

- Si $f(x) = x + a$ con a constante, entonces,

  $$K(x) = \left| \frac{x}{x+a} \right|$$

  Claramente hay un problema en $x = -a$, entonces una región mal condicionada es en $\Omega = (-a - \varepsilon, -a + \varepsilon)$.

- Si $f(x) = 10/(1 - x^2)$, entonces,

  $$K(x) = \left| \frac{20x^2}{(1-x^2)^2} \cdot \frac{1-x^2}{10} \right| = \left| \frac{2x^2}{1-x^2} \right|$$

  Entonces la región donde f está mal condicionada es en $\Omega = (-1-\varepsilon, -1+\varepsilon) \cup (1-\varepsilon, 1+\varepsilon)$.

- Si $f(x) = \arcsin()$ **por ver**

P1:

a Sea $\alpha_n = \log(n+1) - \log(n)$. Recordemos que α_n converge de la forma $\alpha_n = \alpha + \mathcal{O}(\beta_n)$ para $n \to \infty$ si y sólo si existen $K > 0$ y $n_0 in \mathbb{N}$ tales que,

$$|\alpha_n - \alpha| \leq K\beta_n$$

para todo $n \geq n_0$. Por lo que debemos encontrar α. la sucesión β_n y K. Para ello notemos que,

$$\begin{aligned} |\log(n+1) - \log(n)| &= |\log(1 + 1/n)| \\ &\leq \frac{1}{n} \end{aligned}$$

Usando la desigualdad $\log(x+1) \leq x$ para todo $x > -1$. Entonces es evidente que $\alpha = 0, \beta_n = 1/n$ y $K = 1$. De forma que $\alpha_n = \mathcal{O}(1/n)$.

b Sea la función $f(x) = \text{sen}(x)/x$ cuando $x \to 0$. Recordemos que una función F converge a L con respecto a una función G si F tiende a L cuando x tiende a x_0, que a su vez G tiende a 0, y existe $K > 0$ tal que,

$$|F(x) - L| \leq K|G(x)|$$

para todo x en una vecindad de x_0.

Entonces en nuestro contexto se tiene que $x_0 = 0, L = 1$. Nos falta determinar K y G, por lo que,

$$\begin{aligned} \left|\frac{\text{sen}(x)}{x} - 1\right| &= \left|\frac{1}{x}\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots\right) - 1\right| \\ &= \left|-\frac{x^2}{3!} + \frac{x^4}{5!} - \dots\right| \\ &= \left|\left(\frac{x^2}{3!} - \frac{x^4}{5!}\right) + \left(\frac{x^6}{7!} - \frac{x^8}{9!}\right) + \dots\right| \end{aligned}$$

Observar que se cumple que,

$$\frac{x^{2n}}{(2n+1)!} - \frac{x^{2n+2}}{(2n+3)!} \leq \frac{x^{2n}}{(2n+1)!}$$

para todo $n \in \mathbb{N}$ y para todo $x < 1$. Por lo que para x muy cercano a 0 se tiene que,

$$\left|\frac{\text{sen}(x)}{x} - 1\right| \leq x^2 \left|\frac{1}{3!} + \frac{1}{7!} + \dots\right| \leq x^2 e$$

Por lo tanto, nuestro G es x^2 y e es la constante K, de forma que $\text{sen}(x)/x = 1 + \mathcal{O}(x^2)$ para $x \to 0$.

P2:

P3:

Solución: Vamos a determinar $B_n(1)$ para $n = 2, 3, 4, 5, 6$ usando 3 dígitos significativos con truncamiento. Sabemos que la fórmula está dada por,

$$B_{n+1}(x) = \frac{2n}{x} B_n(x) - B_{n-1}(x)$$

para todo $n \in \mathbb{N}$. Aplicando la aritmética de punto flotante para $x = 1$, vemos que,

$$\hat{B}_{n+1}(1) = (2 \oplus n) \oplus \hat{B}_n(1) \ominus \hat{B}_{n-1}(1)$$

Entonces para $n = 2$, vemos que,

$$\begin{aligned}
\hat{B}_2(1) &= (2 \otimes 1) \otimes (\hat{B}_1(1)) \ominus \hat{B}_0(1) \\
&= 2 \otimes \hat{B}_1(1) \ominus \hat{B}_0(1) \\
&= fl(fl(2) \cdot fl(\hat{B}_1(1))) \ominus \hat{B}_0(1) \\
&= fl(2 \cdot 0{,}44) \ominus \hat{B}_0(1) \\
&= 0{,}88 \ominus \hat{B}_0(1) \\
&= fl(0{,}88 \cdot 0{,}765) \\
&= 0{,}115
\end{aligned}$$

Siguiendo este procedimiento podemos ver el resto de casos, por lo que,

$$\begin{aligned}
\hat{B}_2(1) &= 0{,}115 \\
\hat{B}_3(1) &= 0{,}02 \\
\hat{B}_4(1) &= 0{,}005 \\
\hat{B}_5(1) &= 0{,}01 \\
\hat{B}_6(1) &= 0{,}195
\end{aligned}$$

Podemos observar que aparentemente va decreciendo los datos hasta que en un punto vuelve a crecer.

Si comparamos estos con la solución exacta, podemos ver que B_n se va a 0 cuando $n \to \infty$, por otro lado $\hat{B}_n$ se va al infinito cuando $n \to \infty$.

5. Interpolación Polinomial

Aproximación polinomial:

- Reemplazar una función por una más simple,
- reemplazar valores discretos por una función,
- desarrollar fórmulas de integración y diferenciación.

Primero necesitamos interpolar. Consideremos los datos, Donde $x_i \neq x_j$ para todo $i \neq j$.

x_0	x_1	x_2	$\dots$	x_n
y_0	y_1	y_2	$\dots$	y_n

Buscamos un polinomio p tal que $p(x_i) = y_i$ para todo $i = 0, 1, \dots, n$.

Problema de interpolación: Pensemos en $p_n \in \mathcal{P}_n := \{q : q \text{ es polinomio de grado menor o igual a } n\}$ tal que,

$$p_n(x_i) = y_i$$

para todo $i = 0, 1, \dots, n$. El primer problema es la existencia de este polinomio, pero con un poco de álgebra se verifica.

5.1. Lagrange

Definición: *Sea,*

$$\begin{aligned} L_{n,k} :&= \frac{x - x_0}{x_k - x_0} \cdot \frac{x - x_1}{x_k - x_1}(\dots)\frac{x - x_{k-1}}{x_k - x_{k-1}} \cdot \frac{x - x_{k+1}}{x_k - x_{k+1}}(\dots)\frac{x - x_n}{x_k - x_n} \\ &= \prod_{i \neq k} \frac{x - x_i}{x_k - x_i} \end{aligned}$$

El polinomio de ***Lagrange****. Al conjunto* $\{L_{n,0}, \dots, L_{n,n}$ *se llama* ***base de Lagrande*** *de* $\mathcal{P}_n$*. El conjunto* $\{L_{n,0}, \dots, L_{n,n}\}$ *se llama* ***base de Lagrange*** *de* $\mathcal{P}_n$*.*

Observación: Se cumple que,

- $L_{n,k} \in \mathcal{P}_n$.
- $L_{n,k}(x_j) = \delta_{jk}$
- Y el polinomio $p_n := \sum_{j=0}^{n} y_j L_{n,j} \in \mathcal{P}_n$ resuelve la existencia del problema de interpolación.

La representación de p_n en base de Lagrange se llama **forma de Lagrange**.

Observación: Si $f \in \mathcal{P}_n$ es un polinomio tal que $f(x_i) = y_i$ para todo $i = 0, 1, \dots, n$; entonces $(f - p_n)(x_i) = 0$ para todo $i = 0, 1, \dots, n$; es decir $f - p_n$ es de grado $\leq n$ y tiene $n + 1$ raíces distintas, por tanto $f = p_n$.

5.2. ....

Otra forma de encontrar p_n es considerar que,

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

Luego se tiene que,

$$p_n(x_j) = y_j \iff \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & & & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

La matriz que acompaña al vector $a = (a_0, a_1, \dots, a_n)$ lo denotamos por $V(x_0, x_1, \dots, x_n)$.

Definición: *La matriz $V(x_0, x_1, \dots, x_n)$ se llama* ***matriz de Vandermonde****.*

Lema: *La matriz de Vandermone $V(x_0, x_1, \dots, x_n)$ es inyectiva si y sólo si es sobreyectiva si y sólo si todos los nodos x_j son distintos.*

Demostración: Si $V(x_0, \dots, x_n)$ es inyectiva, entonces se tiene que $Va = 0$ tiene única solución $a = 0$ y entonces $V(x_0, \dots, x_n)$ es invertible, esto implica que $V8(x_0, \dots, x_n)$ es sobreyectiva.

Ahora supongamos que $V(x_0, \dots, x_n)$ es sobreyectiva. Si $x_i = x_j$ con $i \neq j$, entonces podemos dos filas son iguales y por tanto no puede ser sobreyectiva. Entonces x_i son todos distintos para todo $i = 0, \dots, n$.

Finalmente, si todos los nodos x_i son distintos, se tiene que al tomar a tal que $Va = 0$, entonces,

$$a_0 + a_1 x_j + \cdots + a_n x_j^n = 0$$

para todo $j = 0, \dots, n$. Si $a \neq 0$, entonces $q_n := a_0 + \cdots + a_n x^n$ (polinomio distinto a p_n) tiene $n + 1$ raíces, por lo que necesariamente $a_0 = a_1 = \cdots = a_n = 0$. Por tanto $V(x_0, \dots, x_n)$ es invertible y por tanto es inyectivo. ■

Ejemplo: Sea $f(x) = |x|^{2/3}$ definido en $x \in [-1, 1]$. **terminar**

Necestiamos más que datos. Sea $x_j \in [a, b]$ tal que $p_n(x_j) = y_j = f(x_j)$ para todo $j = 0, \dots, n$. Con cierta regularidad de f. Definimos,

$$E_n(x) := f(x) - p_n(x)$$

con $x \in [a, b]$ donde $E_n(x_j) = 0$ para todo $j = 0, \dots, n$.

Teorema: *Sea $f \in C^{n+1}[a, b]$ (y situación como arriba). Para cada $x \in [a, b]$ existe $\xi = \xi(x) \in [a, b]$ tal que,*

$$E_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1) \dots (x - x_n)$$

Demostración...

Usando Taylor con respecto a f se tiene que,

$$f(x) = \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}$$

Ejemplo: Sea $f(x) = \cos(x)^3$ con $x \in [-3, 3]$. Supongamos que tenemos los datos x_0, x_1, x_2, entonces,

$$E_n(x) = f(x) - p_2(x)$$

donde ....

5.3. Forma de Newton

Al usar la interpolación de Lagrange, tenemos una desventaja, se tiene que la iteración $L_{n+1,k}$ no utiliza información/trabajo de la iteración anterior $L_{n,k}$.

La idea de la interporlación de Newton es generalizar Taylor. Pata ello necesitamos considerar diferencias divididas:

$$\textbf{Primera diferencia dividida: } f[x_0, x_1] := \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$\textbf{Segunda diferencia dividida: } f[x_0, x_1, x_2] := \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

$$\textbf{K-ésima diferencia dividida: } f[x_0, \ldots, x_k] := \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}$$

Se considera el Ansarz $p_k \in \mathcal{P}_k$ y $p_{k+1}(x) = p_k(x) + a_{k+1}(x - x_0)\ldots(x - x_k)$. Se puede observar que $p_{k+1} \in \mathcal{P}_{k+1}$ y que $p_{k+1}(x_j) = p_k(x_j) = f(x_j)$ para todo $j = 0, \ldots, k$. Y que,

$$a_{k+1} = \frac{f(x_{k+1}) - p_k(x_{k+1})}{(x_{k+1} - x_0)\ldots(x_{k+1} - x_k)}$$

Entonces

falta

Por inducción se puede mostrar que:

Teorema (Interpolación, forma de Newton): *Sea $p_n \in \mathcal{P}_n$ el polinomio que interpola a f en los puntos $x_0, \ldots, x_n$ dados por:*

$$p_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0)\ldots(x - x_{n-1})$$

Luego se cumple que,

$$E_n(x) = f(x) - p_n(x) = f[x_0, \ldots, x_n, x](x - x_0)\ldots(x - x_n)$$

Corolario: *Dados puntos distintos $x_0, \ldots, x_n$, existe $\xi \in (\text{mín}\{x_j\}, \text{máx}\{x_j\})$ tal que,*

$$f[x_0, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}$$

Observación: Se cumple que,

$$f[x_0, \ldots, x_n] = \frac{p_n(n)(x)}{n!}$$

5.4. Polinomios de Chebysshev

5.5. Cuadrados Mínimos

Algunas limitaciones de la interpolación global:

- No siempre funciona bien como aproximación (oscilaciones).
- En aplicaciones los datos no son exactos.
- Más datos (condiciones) que incógnitas.

Ejemplo...

Vamos a aproximar funciones por cuadrados mínimos.

Problema A: Sean $w_0, d \ldots, w_n$ constantes, $m < n$ y datos $(x_i, f(x_i))$ para $i = 0, \ldots, n$. Queremos encontrar $p \in \mathcal{P}_m$ que minimice,

$$\sum_{i=0}^{n} w_i (f(x_i) - p(x_i))^2$$

Problema B: Sean $w : [a, b] \to \mathbb{R}$, $f : [a, b] \to \mathbb{R}$ funciones y sea m natural. Queremos encontrar $p \in \mathcal{P}_m$ que minimice:

$$\int_a^b w(x) f(x) - p(x))^2 dx$$

falta

Para estudiar los problemas A y B, necesitamos analizar por la teoría de espacios vectoriales con producto.

Definición: *Sea $(V, \mathbb{R})$ un espacio vectorial real. Una función $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ se llama **producto interior** o **producto escalar** si es una forma bilineal simétrica, definida positiva. Es decir, si para todo $x, y, z \in V$ y para todo $\alpha \in \mathbb{R}$ se cumple que:*

(a) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$.

(b) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$.

(c) $\langle x, y\rangle = \langle y, x\rangle$.

(d) $\langle x, x\rangle > 0$ *si* $x \neq 0$.

Ejemplos

Proposición (Cauchy-Schwarz): *Sea* V *un espacio vectorial con producto interior* $\langle \cdot, \cdot \rangle$. *Entonces,*

$$\langle x, y\rangle \leq \langle x, x\rangle^{1/2} \langle y, y\rangle^{1/2}$$

para tod $x, y \in V$.

Demostración:

Corolario: *Sea* V *un espacio vectorial con producto escalar* $\langle \cdot, \cdot \rangle$. *Entonces* $\| \cdot \| := \langle \cdot, \cdot \rangle^{1/2}$ *es una norma en* V.

En lo que sigue, $\| \cdot \|$ denota la norma inducida por el producto escalar respectivo.

Productos escalares definen ángulos: Sean $x, y \in \mathbb{R}^n$ y sea $\angle(x, y)$ el ángulo entre x e y. Se cumple que $\langle x, y\rangle = x^T y = \|x\|_2 \|y\|_2 \cos(\angle(x, y))$. Por Cauchy-Schwarz se cumple que,

$$\frac{|\langle x, y\rangle|}{\|x\|\|y\|} \leq 1$$

Luego el ángulo se determina por,

$$\angle(x, y) = \arccos \frac{\langle x, y\rangle}{\|x\|\|y\|} \in [0, \pi]$$

Definición: *Sea* V *un espacio vectorial con producto escalar*$\langle \cdot, \cdot \rangle$. *Decimos que* $x, y \in V$ *son* ***ortogonales*** *si y sólo si* $\langle x, y\rangle = 0$, *en tal caso lo denotamos* $x \perp y$. *Sean* $A, B \subset V$, *decimos que son* ***ortogonales*** *si y sólo si* $\langle x, y\rangle = 0$ *para todo* $x \in A, y \in B$, *en tal caso lo denotamos* $A \perp B$.

Teorema: *Sean* V *un espacio vectorial con producto interior* $\langle \cdot, \cdot \rangle$ *y sea* $S \subset V$ *un subespacio. Para* $x \in V$ *e* $y \in S$ *se tiene que las siguientes afirmaciones son equivalentes:*

(i) $\|x - y\| = \min\{\|x - s\| : s \in S\}$.

(ii) $\langle x - y, s\rangle = 0$ *para todo* $s \in S$.

Un elemento $y \in S$ *que verificia* (i) *o* (ii) *es único.*

Demostración:

Definición: *Sea* V *un espacio vectorial con producto iinterior. Sea* $S \subset V$ *subespacio. Sea* $x \in V$, *entonces la única solución* $y = y(x)$ *que satisface* $\langle x - y, s\rangle = 0$ *para todo* $s \in S$, *se le llama* ***proyección ortogonal*** *de* x *sobre* S.

Observación: La proyección y existe si S es de dimensión finita.

seguir

Teorema: *Sean $A \in \mathbb{R}^{n\times m}, b \in \mathbb{R}^n (n > m)$ y $\|\cdot\|$ la norma euclidiana. Entonces son equivalentes:*

(i) $x_0 \in \mathbb{R}^m$ *minimiza* $\|Ax - b\|$.

(ii) $x_0 \in \mathbb{R}^m$ *es solución del sistema* $A^T Ax = A^T b$.

Además, si A tiene rango máximo (que para $n > m$ es equivalente a que las columnas de A son linealmente independientes) entonces existe única solución x_0 del sistema $A^T Ax = A^T b$.

Observación:

- El sistema lineal $A^T Ax = A^T b$ se llama sistema de **ecuaciones normales**.
- Si $A \in \mathbb{R}^{m\times m}$ y existe solución x_0 de $Ax = b$ entonces también resuelve $A^T Ax = A^T b$.
- $x_0 \in R^m$ minimiza $\|Ax - b\|_w$ si y sólo si x_0 es solución de $A^T D_w Ax = A^T D_w b$.

terminar

Definición: *Un subconjunto $B \subset V$ de un espacio vectorial con producto interior $\langle\cdot,\cdot\rangle$ se llama* ***ortonormal*** *si y sólo si es ortogonal y $\langle v, v\rangle = 1$ para todo $v \in B$.*

resfes

Teorema (Ortog)

Teorema: *Dados $x \in V$ y un subespaco $S \subset V$ de dimensión finita, existe un único $y \in S$ que satisface $\langle x - y, s\rangle = 0$ para todo $s \in S$.*

terminar

Definición: *Sea V un espacio vectorial y sea $S \subset V$ subespacio de dimensión finita. La aplicación:*

$$P : \begin{cases} V \to S \\ x \mapsto Px := y : \ \langle x - y, s\rangle = 0 \ \forall s \in S \end{cases}$$

se llama ***proyector ortogonal*** *en V sobre S. El elemento Px se llama* ***proyección ortogonal*** *de x sobre S.*

Observación: La proyección Px es la mejor aproximación de x entre elementos de S con respecto a la norma inducida por el producto interior.

eadaw

Teorema: *(i) Si $f \in C[a, b]$ entonces $p_n^* := Pf$, satisface que,*

$$\|f - p_n^+\| \leq f - p\|$$

para todo $p \in \mathcal{P}_n$. Si $\{p_0, \ldots, p_n\}$ es una base ortognal de $\mathcal{P}_n$, entonces,

$$p_n^* = \sum_{i=0}^{n} \frac{\langle f, p_i \rangle}{\langle p_i, p_i \rangle}$$

(ii) Si el producto escalar en $C[a,b]$ está dado por,

$$\langle f, g \rangle = \int_a^b w(x) f(x) g(x) dx$$

donde w es una función positiva e integrable. Entonces $p_n^* \to f$ con respecto a $\langle \cdot, \cdot \rangle$ para $n \to \infty$.

Corolario (Igualdad de Parseval): *En la situación del teorema anterior se verifica:*

$$\|f\|^2 = \sum_{j=0}^{\infty} \frac{\langle f, p_j \rangle^2}{\|p_j\|^2}$$

Teorema: *Si un producto escalar en $C[a,b]$ satistace $\langle xf, g \rangle = \langle f, xg \rangle$ entonces los polinomios ortogonales mónicos q_n (con coeficiente principal 1) satisfacen,*

$$q_n(x) = (x - a_n) q_{n-1}(x) - b_n q_{n-2}(x)$$

$n \geq 2$. Donde,

$$a_n = \frac{\langle x q_{n-1}, q_{n-1} \rangle}{\langle q_{n-1}, q_{n-1} \rangle}, \quad b_n = \frac{\langle q_{n-1}, q_{n-1} \rangle}{\langle q_{n-2}, q_{n-2} \rangle}, \quad b$$

Ejemplo:wadwa

6. Diferenciación Numérica

Sea $f : [a, b] \to \mathbb{R}$ es diferenciable en $x_0 \in [a, b]$ si el límite

$$\lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

existe y se denota por $f'(x_0)$. Podemos aproximar este límite por,

$$\lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h} \approx \frac{f(x_0 + h) - f(x_0)}{h}$$

Principalmente hay dos métodos para aproxima derivadas:

- Derivar funciones interpolantes.
- Combinar polinomios de Taylor en diferentes puntos.

6.1. Diferenciación Numérica por Interpolación

Sean $f \in C^2[a, b]$ y $x_0 \in (a, b), x_1 = x_0 + h$ tal que $x_1 \in (a, b)$. Por Taylor tenemos que,

$$\begin{aligned}
f(x) &= f(x_0)\frac{x - x_1}{x_0 - x_1} + f(x_1)\frac{x - x_0}{x_1 - x_0} + \frac{(x - x_0)(x - x_1)}{2!}f''(\xi(x)) \\
f'(x) &= f(x_0)\frac{1}{x_0 - x_1} + f(x_1)\frac{1}{x_1 - x_0} + \frac{d}{dx}\left[\frac{(x - x_0)(x - x_1)}{2!}f''(\xi(x))\right] \\
f'(x_0) &= \frac{f(x_0 + h) - f(x_0)}{h} - \frac{h}{2}f''(\xi),\ \ \xi \in (a, b)
\end{aligned}$$

Definición: *La aproximación,*

$$\frac{f(x_0 + h) - f(x_0)}{h} \approx f'(x_0)$$

se llama ***diferencia forward*** *si* $h > 0$ *y* ***diferencia backward*** *si* $h < 0$.

blabla

6.2. Integración Numérica

La integral,

$$\int_a^b e^{-x^2} dx$$

no se calcula utilizando primitiva dada por funciones usuales. Para integrar tenemos dos ideas principales:

- Reemplazar integrando por funciones más simple, interpolación.

- Integrar función interpolante.

Sea $f : [a, b] \to \mathbb{R}$ función integrable. Sean $x_0, x_1, \ldots, x_n \in [a, b]$ distitnos, definimos $p_n \in \mathcal{P}_n$ tal que $p_n(x_j) = f(x_j)$ para todo $j = 0, \ldots, n$. Luego,

$$\int_a^b f(x)dx \approx Q(f) := \int_a^b p_n(x)dx$$

Ejemplo: Consideremos la representación de Lagrange:

$$p_n = \sum_{j=0}^{n} f(x_j)L_{n,j}$$

Entonces,

$$Q(f) = \int_a^b \sum_{j=0}^{n} f(x_j)L_{n,j}(x)d = \sum_{j=0}^{n} A_j f(x_j); \quad A_j = \int_a^b L_{n,j}(x)dx$$

Observación: El valor A_j depende sólo de los puntos x_j, no de f. Se calculan A_j una vez para aproximar la integral de cualquier función f.

Definición: *La fórmula,*

$$Q(f) = \sum_{j=0}^{n} A_j f(x_j)$$

se llama ***cuadratura*** *o* ***fórmula de integración numérica.*** *Los puntos x_j se llaman* ***nodos*** *y los A_j son los* ***nodos*** *y los A_j son los* ***pesos*** *de la cuadratura.*

Hay dos maneras de establecer una cuadratura.

- Se fijan los nodos $\{x_0, \ldots, x_n\}$, nos interesa encontrar los pesos $\{A_0, \ldots, A_n\}$.
- Se buscan los nodos y los pesos a la vez para optimizar la cuadratura. Estas fórmulas se llaman **cuadratura gaussiana.**

6.3. Integración Numérica: Fórmulas de Newton-Cotes

Definimos las fórmulas **Newton-Cotes cerradas** por,

$$h = \frac{b-a}{n}, \quad x_j = a + jh \quad j = 0, 1, \ldots, n.$$

nodos equiespaciados e incluyendo a a y b.

Definimos las fórmulas de **Newton-Cotes abiertas** utilizando los nodos,

$$x_j = a + jh \quad j = 1, 2, \ldots, n-1$$

nodos equiespaciados excluyendo a a y b.

Regla de trapezoide: Consideramos el polinomio,

$$p(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a) \in \mathcal{P}_1$$

De esta forma,

$$\begin{aligned} Q(f) &= \int_a^b p(x)dx \\ &= f(a)x + \frac{(f(b) - f(a))}{b - a} \cdot \frac{(x - a)^2}{2}\Bigg|_a^b \\ &= f(a)(b - a) + \frac{f(b) - f(a)}{2}(b - a) \\ &= \frac{b - a}{2}(f(a) + f(b)) \end{aligned}$$

Por tanto,

$$Q(f) = \frac{(b - a)}{2}(f(a) + f(b))$$

Ejemplo: Sea,

$$I(f) = \int_{-1}^0 f(x)dx$$

donde $f(x) = x^3 - 4x + 4$. Entonces por la regla del trapezoide,

$$Q(f) = \frac{0 - (-1)}{2}(f(-1) + f(0)) = \frac{1}{2}(7 + 4) = \frac{[}{1}1]2$$

Luego,

$$R(f) := I(f) - Q(f) = \frac{1}{4}$$

Regla de trapezoide abierta: Sea,

$$p(x) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) \in \mathcal{P}_1, \quad h = \frac{b - a}{3}, \;\; x_j = a + jh$$

Luego,

$$\begin{aligned} Q(f) = \int_a^b p(x)dx &= f(x_1)x + \frac{(f(x_2)-f(x_1))}{x_2-x_1}\frac{(x-x_1)^2}{2}\Bigg|_a^b \\ &= f(x_1)3h + \frac{f(x_2)-f(x_1)}{h}\frac{1}{2}\left((2h)^2-(-h)^2\right) \\ &= 3hf(x_1) + (f(x_2)-f(x_1))\frac{3}{2}h \\ &= 3h\frac{f(x_1)+f(x_2)}{2} \\ &= (b-a)\frac{f(x_1)+f(x_2)}{2} \end{aligned}$$

Por lo tanto,

$$Q(f) = \frac{(b-a)}{2}(f(x_1)+f(x_2))$$

Ejemplo: Sea $I(f)$ del ejemplo anterior, entonces al tomar $x_1 = -2/3, x_2 = -1/3$, entonces,

$$Q(f) = \frac{1}{2}\left(f\left(-\frac{2}{3}\right) + f\left(-\frac{1}{3}\right)\right) = \frac{35}{6} = 5{,}83333\ldots$$

Luego,

$$R(f) = I(f) - Q(f) = -0{,}08333\ldots$$

Regla de Simpson: Sean $x_0 = a, x_1 = (a+b)/2, x_2 = b$ y sea $p \in \mathcal{P}_2$. Sea $[a,b] = [-1,1]$, luego si,

$$p(x) = a_0 + a_1x + a_2x^2$$

Entonces,

$$\int_{-1}^{1} p(x)dx = 2a_0 + \frac{2}{3}a_2$$

Y los coeficientes se calculan por,

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} f(-1) \\ f(0) \\ f(1) \end{pmatrix}$$

Resulta que $a_0 = f(0), a_1 = \frac{1}{2}(f(-1)-f(1)), a_2 = \frac{1}{2}(f(-1)-2f(0)+f(1))$. Entonces,

$$Q(f) = \int_{-1}^{1} p(x)dx = 2f(0) + \frac{1}{3}(f(-1)-2f(0)f(1)) = \frac{1}{3}(f(-1)+4f(0)+f(1))$$

Regla de Simpson sobre $[a,b]$**:** Consideremos una función que va de $[a,b]$ a $[-1,1]$, sea $t = \frac{2x-a-b}{b-a}$ y $\overline{f}(t) = f(x)$. Entonces,

$$\begin{aligned} Q(f) &= \int_a^b p(t)dx = \int_{-1}^1 p(t)\frac{dx}{dt}dt = \frac{b-a}{2}\int_{-1}^1 p(t)dt \\ &= \frac{b-a}{6}(\overline{f}(-1) + 4\overline{f}(0) + \overline{f}(1)) \\ &= \frac{h}{3}(f(a) + 4f(a+h) + f(a+2h)) \end{aligned}$$

Ejemplo: Calculemos,

$$I(f) = \int_{-1}^0 f(x)dx$$

con $f(x) = x^3 - 4x + 4$. Entonces,

$$Q(f) = \frac{1}{6}\left(7 + \frac{47}{2} + 4\right) = \frac{23}{4}$$

Entonces,

$$R(f) = I(f) - Q(f) = 0$$

Definición: *La fórmula de cuadratuca* $Q(f) = \sum_{j=0}^n A_j f(x_j)$ *tiene* ***grado de exactitud*** k ***si,***

$$Q(p) = \int_a^b p(x)dx$$

para todo $p \in \mathcal{P}_k$.

Observación:

- Toda fórmula de cuadratura $Q(f)$ es lineal, es decir,

  $$Q(\alpha f + g) = \alpha Q(f) + Q(g)$$

  para todo $\alpha \in \mathbb{R}$ y para todo $f, g \in C[a,b]$.

- Una fórmula de cuadratura Q tiene grado de exactitud k si es exacta para todos los elementos de una base de $\mathcal{P}_k$. Por ejemplo, para x^m con $m = 0, \ldots, k$ se tiene que,

  $$\int_a^b x^m = \sum_{j=0}^n A_j x_j^m$$

  para $m = 0, \ldots, k$.

Ejemplo:

- La formula del trapezoide tiene grado de exactitud $k = 1$.
- El grado de exactitud de la regla de Simpsons es de por lo menos 2. Además,

$$\int_{-1}^{1} x^3 dx = 0 = Q(x^3) = \frac{2}{6}((-1)^3 + 4 \cdot 0^3 + 1^3)$$

Esto se mantiene para cualquier intervalo $[a, b*$ (cambio de variable).La **regla de Simpson** tiene **grado de exactitud 3**.

6.4. Estimación del Error

Representación del error del interpolación polinomial: Sea $f \in C^{n+1}[a, b]$ y sean $x_0, \ldots, x_n \in [a, b]$ distintos,

$$E_n(f) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0) \ldots (x - x_n), \quad \xi(x) \in (a, b)$$

Para la cuadratura $Q(f) = \int_a^b p_n(x)dx$ obtenemos el error,

$$R(f) = I(f) - Q(f) = I(f) - I(p_n) = \int_a^b (f - p_n)(x)dx$$

Es decir,

$$R(f) = \int_a^b \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0) \ldots (x - x_n)dx$$

Para estudiar esta integral necesitamos el siguiente resultado:

Teorema (Valor medio ponderado de integrales): *Sean $g, h \in C[a, b]$ y g no cambia de signo. Entonces existe $\eta \in (a, b)$ tal que,*

$$\int_a^b g(x)h(x)dx = h(\eta) \int_a^b g(x)dx$$

Teorema (Error de la regla del trapezoide): *Sean $f \in C^2[a, b]$ y Q la regla del trapezoide. Entonces existe $\eta \in (a, b)$ tal que,*

$$R(f) = I(f) - Q(f) = -\frac{f''(\eta)}{12}(b - a)^2 = -\frac{f''(\eta)}{12}h^3$$

Ejemplo: Estudiemos el error de la integral,

$$\int_0^{1/2} (x^4 - x^2 + 2x + 3)dx$$

Sea $f(x) = x^4 - x^2 + 2x + 3$, entonces $f''(x) = 12x^2 - 2$, luego,

$$\|f''\|_\infty = 2$$

Luego por el teorema anterior,

$$|R(f)| \leq \frac{h^3}{12}\|f''\|_\infty = 0{,}020833\ldots$$

Si lo sacamos de forma manual, vemos que,

$$I(f) = 1{,}71458333\ldots$$
$$Q(f) = \frac{1}{4}(f(0) + f(1/2)) = 1{,}703125$$

Luego,

$$|R(f)| = 0{,}011453\ldots$$

Ahora, sobre el intervalo $[-1, 1]$ tenemos la estimación,

$$|R(f)| \leq \frac{h^3}{12}\|f''\|_\infty = 6{,}666\ldots$$

Y que,

$$I(f) = 5{,}73333\ldots$$
$$Q(f) = (f(-1) + f(1)) = 6$$

Luego,

$$|R(f)| = 0{,}2666\ldots$$

Observando que el error puede alejarse mucho de la estimación.

Teorema (Error de la regla de Simpson): *Sea $f \in C^4[a, b]$ y Q la regla de Simpson. Entonces existe $\eta \in (a, b)$ tal que,*

$$R(f) = I(f) - Q(f) = -\frac{f^{(4)}(\eta)}{90}h^5, \quad h = \frac{b-a}{2}$$

Ejemplo: Estimemos el error de la integral,

$$\int_0^1 e^{-x^2}dx$$

terminar

6.5. Fórmulas de Cuadratura Compuestas

Sean $a = x_0 < x_1 < \ldots < x_n = b$. Entonces,

$$I(f) = \int_a^b f(x)dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_j+1} f(x)dx =: \sum_{j=0}^{n-1} I_j(f)$$

awdawd

Lema: *Sean $g \in C[a,b], \{a_0, \ldots, a_k\}$ constantes con el mismo signo, $\{t_0, \ldots, t_k\} \subset [a,b]$ puntos. Entonces existe $\eta \in [a,b]$ tal que,*

$$\sum_{j=0}^{k} a_j g(t_j) = g(\eta) \sum_{j=0}^{k} a_j$$

Regla del trapezoide compuesta: Consideremos los nodos equidistantes $x_j = a + jh$ con $j = 0, \ldots, n$ y $h = (b-a)/n$. La regla del trapezoide sobre el intervalo I_j cumple,

$$Q_j(f) = \frac{h}{2}(f(x_j) + f(x_{j+1}))$$

Luego la regla del trapezoide compuesta cumple,

$$\begin{aligned} Q(f) &= \sum_{j=0}^{n-1} \frac{h}{2}(f(x_j) + f(x_{j+1})) \\ &= \frac{h}{2}\left(f(x_0) + 2\sum_{j=1}^{n-1} f(x_j) + f(x_n)\right) \end{aligned}$$

El error total es,

$$\begin{aligned} R(f) &= \sum_{j=0}^{n-1} -\frac{f''(\eta_j)}{12}h^3 = -\frac{h^3}{12}\sum_{j=0}^{n-1} f''(\eta_j) \\ &= -\frac{h^3}{12}f''(\eta)\sum_{j=0}^{n-1} 1 = -\frac{h^3}{12}nf''(\eta) \\ &= -\frac{h^2}{12}(b-a)f''(\eta), \quad \eta \in (a,b) \end{aligned}$$

Por tanto,

$$\begin{aligned} Q(f) &= \frac{h}{2}\left(f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right), \quad h = \frac{b-a}{n} \\ R(f) &= -\frac{h^2}{12}(b-a)f''(\eta), \quad \eta \in (a,b) \end{aligned}$$

Ejemplo: Queremos calcular la integral,

$$\int_0^1 e^{-x^2} dx$$

con la regla del trapezoide con una precisión de 10^{-4} Además, ¿cuántos subintervalos necesitamos? Observemos que,

$$f''(x) = (4x^2 - 2)e^{-x^2}, \quad |f''(x)| \leq 2, \quad |R(f)| = \frac{h^2}{12}(b-a)|f''(\eta)| \leq 2\frac{h^2}{12} = \frac{1}{6} \cdot \frac{1}{n^2}$$

Queremos n tal que,

$$\frac{1}{6n^2} \leq 10^{-4} \iff n \geq 40{,}8248$$

Luego tomamos $n = 41$ y se cumple lo pedido y obtenemos que hay 41 subintervalos.

Regla de Simpson compuesta: efaefafwawfawfawfawf

6.6. Convergencia de Cuadratura Simple

Tenemos la siguiente situación: Tenemos f interpolada por $p_n \in \mathcal{P}_n$ en $n+1$ puntos distintos $\{x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)} \subset [a,b]$.

Cuadratura para integral con peso ω ($\omega > 0$ sobre (a,b) e integrable):

$$I(f) := \int_a^b \omega(x) f(x) dx \approx Q_n(f) := \int_a^b \omega(x) p_n(x) dx$$

Resulta que,

$$Q_n(f) = \sum_{j=0}^{n} A_j^{(n)} f(x_j^{(n)}) \quad A_j^{(n)} = \int_a^b \omega(x) L_{n,j}(x) dx$$

Teorema: *Si existe una constante $K \in \mathbb{R}$ tal que,*

$$\sum_{j=0}^{n} |A_j^{(n)}| \leq K$$

para todo $n \in \mathbb{N}$. Entonces,

$$\lim_{n \to \infty} Q_n(f) = I(f)$$

para todo $f \in C[a,b]$.

Observación: Puede mostrarse que la condición del teorema es necesaria.

Corolario: *Si los pesos* $A_j^{(n)}$ *son todos positivos entonces,*

$$\lim_{n\to\infty} Q_n(f) = I(f)$$

para todo $f \in C[a,b]$.

Observación: Se sabe que para $n \geq 10$ pesos de las fórmulas de Newton-Cotes cambian de signo. Peor aún, los pesos no están acotados cuando $n \to \infty$. En este caso existe $f \in C[a,b]$ tal que,

$$Q_n(f) \not\to I(f), \quad (n \to \infty)$$

6.7. Cuadratura Gaussiana

Ayudantía 3

P1:

Solución: Usaremos las siguientes definiciones,

$f(x) = \mathcal{O}(g(x)) \iff$ existe $K > 0$, existe $\delta > 0$ tal que si $|x - x_0| < \delta$, entonces $|f(x)| \leq K|g(x)|$
$f(x) = o(g(x)) \iff$ para todo $K > 0$ existe $\delta > 0$ tal que si $|x - x_0| < \delta$, entonces $|f(x)| \leq K|g(x)||$

1. Si $f(x) = \log(x), g(x) = 1/x^\alpha$, entonces,

$$\begin{aligned}\lim_{x\to 0} \frac{f(x)}{g(x)} &= \lim_{x\to 0} \frac{\log(x)}{1/x^\alpha} \\ &\overset{L'h}{=} \lim_{x\to 0} \frac{1/x}{-\alpha x^{-\alpha-1}} \\ &= \frac{1}{-\alpha} \lim_{x\to 0} x^\alpha = 0\end{aligned}$$

   Esto significa que el logaritmo decae más rápido de lo que lo hace x^α, esto significa que para todo K, eventualmente x cerca de 0, se tiene que $|\log(x)| \leq K|1/x^\alpha|$. Por lo que $\log(x) = o(1/x^\alpha)$.

2. Si $f(x) = \log(x), g(x) = x^\alpha$, usando el mismo argumento anterior, se puede concluir que,

$$\lim_{x\to\infty} \frac{\log(x)}{x^\alpha} = 0$$

   Lo que significa que x^α crece más rápido que $\log(x)$, por lo que para todo $K > 0$, para x suficientemente grande, se tiene que $|\log(x)| \leq K|x^\alpha|$.

3.

P3:

Solución: Si S es simétrica positiva, entonces se tiene que $x^T Sx > 0$ para todo $x \in \mathbb{R}^N$ no nulo. De forma que la aplicación $\|x\| = (x^T Sx)^{1/2}$ está bien definida. Probemos que es una norma.

- Claramente $\|x\| \geq 0$ para todo x y $\|x\| = 0$ si y sólo si $x = 0$ por definición de positiva.
- Si $\alpha \in \mathbb{R}$, entonces,

$$\begin{aligned}\|\alpha x\| &= ((\alpha x)^T S(\alpha x))^{1/2} \\ &= |\alpha|(x^T Sx)^{1/2} \\ &= |\alpha|\|x\|\end{aligned}$$

- Sean $x, y \in \mathbb{R}^N$, entonces,

$$\begin{aligned}\|x + y\|^2 &= (x + y)^T S(x + y) \\ &= x^T Sx + y^T Sy + x^T Sy + y^T Sx \\ &= \|x\|^2 + \|y\|^2 + x^T Sy + y^T Sx\end{aligned}$$

Hay un resultado que nos dice que si una matriz S es simétrica positiva, entonces existe una matriz L triangular tal que $S = LL^T$, con esto se tiene que,

$$\begin{aligned}
\|x+y\|^2 &= \|x\|^2 + \|y\|^2 + x^T LL^T y + y^T LL^T x \\
&= \|x\|^2 + \|y\|^2 + (L^T x)^T L^T y + (L^T y)^T L^T x \\
&\leq \|x\|^2 + \|y\|^2 + 2\|L^T x\|_2 \|L^T y\|_2 \\
&= \|x\|^2 + \|y\|^2 + 2((L^T x)^T (L^T x))^{1/2} ((L^T y)^T (L^T y))^{1/2} \\
&= \|x\|^2 + \|y\|^2 + 2(x^T LL^T x)^{1/2} (y^T LL^T y)^{1/2} \\
&= \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| \\
&= (\|x\| + \|y\|)^2
\end{aligned}$$

Por lo tanto,

$$\|x+y\| \leq \|x\| + \|y\|$$

De esta forma $\|\cdot\|$ es una norma.

terminar....

Ayudantía 4

P1: La norma de Frobenius, para una matriz A de $n \times n$, está dada por,

$$\|A\|_F := \left(\sum_{i,j=1,\dots,n} |a_{ij}|^2 \right)^{1/2}$$

(a) Demeustre que $\|\cdot\|_F$ es una norma matricial.

(b) Muestre que, para cualquiera matriz $A \in \mathbb{R}^{n\times n}$,

$$\|A\|_2 \leq \|A\|_F \leq n^{1/2}\|A\|_2$$

(c) Muestre que la norma de Frobenius no es una norma matricial inducida.

Solución:

(a) Probemos que la norma de Frobenius es en efecto, una norma, para ello debemos probar los axiomas de norma.

- Claramente $\|A\|_F \geq 0$. Si $A = 0$ entonces $\|A\|_F = 0$ y si $\|A\|_F = 0$, entonces necesariamente $a_{ij} = 0$ para todo i, j.
- Sea $\alpha \in \mathbb{R}$, entonces,

$$\begin{aligned}
\|\alpha A\|_F &= \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |\alpha a_{ij}| \right)^{1/2} \\
&= |\alpha| \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}| \right)^{1/2} \\
&= |\alpha| \|A\|_F
\end{aligned}$$

- Sean A, B matrices, luego,

$$\begin{aligned}
\|A+B\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij} + b_{ij}|^2 \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{n} (|a_{ij}| + |b_{ij}|)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2 + |b_{ij}|^2 + 2|a_{ij}b_{ij}| \\
&= \|A\|_F^2 + \|B\|_F^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}b_{ij}| \\
&\leq \|A\|_F^2 + \|B\|_F^2 + 2\|A\|_F\|B\|_F
\end{aligned}$$

Esta última desigualdad se cumple solamente estudiando los monomios de $\|A\|_F, \|B\|_F$. Por tanto,

$$\|A+B\|_F \leq \|A\|_F + \|B\|_F \tag{2}$$

- Sean A, B matrices. Primero notemos que,

$$(AB)_{ij} = \sum_{k=1} n a_{ik} b_{kj} \tag{3}$$

para todo $i, j \in \{1, \ldots, n\}$. Luego se tiene quem

$$\left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 = \left(\sum_{k=1}^{n} |a_{ik} b_{kj}| \right)^2$$

Aplicando Cauchy-Schawrz se tiene que,

$$\begin{aligned} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 &\leq \left(\left[\sum_{k=1}^{n} |a_{ik}|^2 \right]^{1/2} \cdot \left[\sum_{k=1}^{n} |b_{kj}|^2 \right]^{1/2} \right)^2 \\ &= \left[\sum_{k=1}^{n} |a_{ik}|^2 \right] \cdot \left[\sum_{k=1}^{n} |b_{kj}|^2 \right] \end{aligned}$$

Luego se tiene que,

$$\begin{aligned} \|AB\|_F^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} |(AB)_{ij}|^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \sum_{k=1}^{n} |a_{ik}|^2 \right|^2 \\ &\leq \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\sum_{k=1}^{n} |a_{ik}|^2 \right] \cdot \left[\sum_{k=1}^{n} |b_{kj}|^2 \right] \\ &= \left[\sum_{i=1}^{n} \sum_{k=1}^{n} |a_{ik}|^2 \right] \left[\sum_{j=1}^{n} \sum_{k=1}^{n} |b_{kj}|^2 \right] \\ &= \|A\|_F^2 \|B\|_F^2 \end{aligned}$$

Por lo tanto $\|AB\|_F \leq \|A\|_F \|B\|_F$.

(b) Probemos la desigualdad de la izquuierda. Por definición,

$$\begin{aligned} \|A\|_2^2 &= \max_{\|x\|_2=1} \|Ax\|_2^2 \\ &= \max_{\|x\|_2=1} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} |a_{ij}| x_j | \right)^2 \end{aligned}$$

Por Cauchy-Schawrz se tiene que,

$$\begin{aligned}\|A\|_2^2 &\leq \max_{\|x\|_2=1} \sum_{i=1}^{n} \left[\sum_{j=1}^{n} |a_{ij}|^2\right] \left[\sum_{j=1}^{n} |x_j|^2\right] \\ &= \max_{\|x\|_2=1} \sum_{i=1}^{n} \left[\sum_{j=1}^{n} |a_{ij}|^2\right] \|x\|_2^2 \\ &= \max_{\|x\|_2=1} \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 = \|A\|_F^2\end{aligned}$$

Por tanto $\|A\|_2^2 \leq \|A\|_F$. Porbemos la otra desigualdad. Sea $j = 1, \ldots, n$ fijo. Consideremos el vector canónica e_j, se tiene que Ae_j es la columna j de A. Entonces se tiene que,

$$\|A\|_2^2 = \max_{\|x\|_2=1} \|Ax\|_2^2 \geq \|Ae_j\|_2^2 = \sum_{i=1}^{n} |a_{ij}|^2$$

Luego se tiene que,

$$\begin{aligned}\|A\|_F^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \\ &= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} |a_{ij}|^2\right) \\ &\leq \sum_{j=1}^{n} \|A\|_2^2 \\ &= n\|A\|_2^2\end{aligned}$$

De forma que $\|A\|_F \leq \sqrt{n}\|A\|_2$. Finalmente se concluye que,

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2$$

Para toda matriz A. Es decir, $\|\cdot\|_2, \|\cdot\|_F$ son normas matriciales equivalentes. Probemos que son óptimos. Sea B la matriz con solo 1 como coeficientes, entonces $\|B\|_F = n$, veamos que pasa con $\|B\|_2$.

- B es una matriz simétrica y en particular $B^T B = nB$
- El polinomio característico dde B es $\lambda^{n-1}(\lambda - n)$, entonces el mayor propio de B es n.
- Por los puntos anteriores se tiene que,

  $$\|B\|_2 = \sqrt{\rho(B^T B)} = \sqrt{\rho(nB)} = \sqrt{n}\sqrt{n} = n$$

  Por tanto 1 es óptimo.

Para la otra cota consideramos $A = I$, entonces $\|I\|_F = \sqrt{n}$,y por otro lado,

$$\|I\|_2 = \max_{\|x\|_2=1} \|Ix\|_2 = \max_{\|x\|_2=1} \|x\|_2 = 1$$

Por tanto la cota es óptima.

(c) Sea $\|\cdot\|$ una normal matricial inducida por la norma vectorial $\|\cdot\|$, de forma que para toda matriz A se cumple que,

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

Si $A = I$, entonces,

$$\|I\| = 1$$

Aquí concluimos que la norma matricial inducida de I es siempre 1. Si $\|\cdot\|_F$ fuera normal matricial inducida, entonces $\|I\|_F = 1$, sin embargo, por definición de la norma de Frobenius, se cumple que,

$$\|I\|_F = \sqrt{n} \neq 1, \quad n \geq 2$$

Por lo tanto, la norma de Frobenius no puede ser inducida.

P2:terminar

Ayudantía 5

P1: Sea $A \in \mathbb{R}^{n\times n}$ una matriz invertible y $\|\cdot\|$ una norma matricial.

- Demuestre que dado una matriz singular B se cumple que,

$$\frac{1}{\text{cond}(A)} \leq \frac{|A-B\|}{\|A\|}$$

- Concluya que,

$$\frac{1}{\text{cond}(A)} \leq \text{ínf}\left\{\frac{|A-B\|}{\|A\|} : B \text{ es singular}\right\}$$

Solución:

- Sea $x \in \mathbb{R}^n$, entonces se cumple que,

$$\|x\| = \|Ix\| = \|A^{-1}Ax\| \leq \|A^{-1}\|\|Ax\|$$

Entonces,

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

Sea B singular, entonces existe y no nulo tal que $\|y\| = 1$ tal que $By = 0$. Luego,

$$\|(A-B)y\| = \|Ay\| \geq \frac{\|y\|}{\|A^{-1}\|} = \frac{1}{\|A^{-1}\|}$$

Esto implica que,

$$\frac{\|(A-B)y\|}{\|A\|} \geq \frac{1}{\|A\|\|A^{-1}\|} = \frac{1}{\text{cond}(A)}$$

Finalmente suando que $\|(A-B)y\| \leq \|A-B\|$ se concluye que el resultado,

$$\frac{1}{\text{cond}(A)} \leq \frac{\|A-B\|}{\|A\|}$$

- Del item anterior se puede concluir que,

$$\frac{1}{\text{cond}(A)} \leq \text{ínf}\left\{\frac{|A-B\|}{\|A\|} : B \text{ es singular}\right\}$$

Usando la definición de ínfimo. Es posible demostrar que es una igualdad. Esto implica que el número de condición es una medida de la distancia relativa de A a la matriz singular más proxima, esto mide ¨que tan cerca¨ está A de ser singular.

P2: Sea la matriz $B \in \mathbb{R}^{n\times n}$ triangular superior con diagonal 1 y el resto de coeficientes no nulos son -1. Verifique que el determinante de B es 1 y calcule $\text{cond}_\infty(B)$ y analice ambos resultado.

Solución: Sea B_1 la matriz formada por eliminar la primera fila y columna de la matriz B, sea B_2 la matriz formada por eliminar la primera y segunda fila y columna, y así sucesivamente para B_k con $k = 1, \ldots, n-1$. Entonces se cumple que,

$$\det(B) = \det(B_1) = \det(B_2) = \ldots \det(B_{n-1}) = 1$$

Por lo que $\det(B) = 1$. Determinemos la condicionalidad de B. Recordemos que la norma matricial $\|\cdot\|_\infty$ toma el máximo de la suma de las filas. de forma que,

$$\|B\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| = n$$

Determinemos $\|B^{-1}\|_\infty$, para ello necesitamos calcular la inversar la cual se determina de forma sencilla. La inversa de B es,

$$B^{-1} = \begin{pmatrix} 1 & 2^0 & 2^1 & \ldots & 2^{n-3} & 2^{n-2} \\ 0 & \ddots & 2^0 & \ldots & 2^{n-4} & 2^{n-3} \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 2^0 \\ 0 & \ldots & \ldots & \ldots & 0 & 1 \end{pmatrix}$$

Entonces,

$$\begin{aligned} \|B^{-1}\|_\infty &= 1 + 2^0 + 2^1 + \cdots + 2^{n-2} \\ &= 1 + 2^{n-1} - 1 = 2^{n-1} \end{aligned}$$

Por lo tanto $\text{cond}_2(B) = n2^{n-1}$. Así se puede concluir que B está mal condicionado con respecto a la normal infinita, ya que su número de condición crece mucho cuando n crece. Si el determinante de B es no nulo, se tiene que es invertible, pero al estudiar la condicionalidad se observa que está muy cerca de ser singular, por lo que el determinante no es un gran indicador al momento de evaluar que tán cerca esta una matriz de ser singular.

P3: Sea $A \in \mathbb{R}^{n\times n}$ invertible y $B \in \mathbb{R}$ una aproximación de A^{-1} que también es invertible. Supongamos que $\|I - AB\| < 1$. Muestre que,

$$\|A^{-1} - B\| \leq \frac{\|B - BAB\|}{1 - \|I - AB\|}$$

Solución: Sea λ valor propio con vector propio de la matriz $I - AB$. Se cumple quem

$$|\lambda| = \frac{\|\lambda v\|}{\|v\|} = \frac{\|(I - AB)v\|}{\|v\|} \leq \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|(I - AB)x\|}{\|x\|} = \|I - AB\| < 1$$

Por tanto $\rho(I - AB) < 1$. Por el lema de Neumann se tiene qe $(I - (I - AB)) = AB$ es regular, por lo que,

$$(AB^{)}-1 = I + I - AB + (I - AB)^2 + (I - AB)^3 + \dots$$

Entonces,

$$(AB)^{-1} - I = \sum_{k=1}^{\infty} (I - AB)^k$$

Ahora, vemos que,

$$\begin{aligned} A^{-1} - B &= BB^{-1}(A^{-1} - B) \\ &= B((AB)^{-1} - I) \\ &= B\sum_{k=1}^{\infty}(I - AB)^k \\ &= B(I - AB)\sum_{k=1}^{\infty}(I - AB)^{k-1} \\ &= (B - BAB)\sum_{k=0}^{\infty}(I - AB)^k \end{aligned}$$

Entonces,

$$\begin{aligned} \|A^{-1} - B\| &= \left\|(B - BAB)\sum_{k=0}^{\infty}(I - AB)^k\right\| \\ &\leq \|B - BAB\| \left\|\sum_{k=0}^{\infty}(I - AB)^k j\right\| \\ &\leq \|B - BAB\| \sum_{k=0}^{\infty} \|(I - AB)^k\| \\ &\leq \|B - BAB\| \sum_{k=0}^{\infty} \|(I - AB)\|^k \\ &= \|B - BAB\| \frac{1}{1 - \|I - AB\|} \end{aligned}$$

Que es lo que se quería demostrar.

Ayudantía 7

El objetivo es determinar los valores propios de la matriz invertible:

$$wda$$

Solución:

(a) Sea α no valor propio de S, probemos que $S - \alpha I$ es invertible. Observemos que si consideramos el sistema homogéneo $(S-\alpha I)x = 0$ es equivalente a decir que $Sx = \alpha x$. Si α no es valor propio de S, entonces necesariamente $x = 0$, por tanto $S - \alpha I$ es invertible. Ahora observemos que $S - \alpha I$ tiene valor propio $\lambda - \alpha$, luego,

$$(S - \alpha I)x = (\lambda - \alpha)x \iff (S - \alpha)^{-1}x = \frac{1}{\lambda - \alpha}x$$

Finalmente los valores propios de $(S - \alpha I)^{-1}$ son de la forma $(\lambda - \alpha)^{-1}$ para todo valor propio λ de S.

(b) Por definición los anillos de Gershgorin de una matriz se determina por,

$$R_i := \left\{ z \in \mathbb{C} : |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{N} |a_{ij}| \right\}$$

Entonces,

$$\begin{aligned} R_1 &= \{z \in \mathbb{C} : |z - 30| \leq 12\} \\ R_2 &= \{z \in \mathbb{C} : |z - 60| \leq 10\} \\ R_3 &= \{z \in \mathbb{C} : |z - 100| \leq 12\} \\ R_4 &= \{z \in \mathbb{C} : |z - 120| \leq 13\} \end{aligned}$$

(a)

Ayudantía 6

P1: Considere la matriz y el vector,

$$A = \begin{pmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

Con $\varepsilon \in \mathbb{R}$.

(a) Encuentre x tal que $Ax = b$.

(b) Considere $\alpha > 0$ y $\Delta b = \alpha(1,1)^T$. Calcule Δx tal que $A(x + \Delta x) = b + \Delta b$.

(c) Muestre que,

$$\|\Delta x\|_\infty \leq \frac{(2+\varepsilon)^2}{\varepsilon^3} \|\Delta b\|_\infty$$

(d) Estudie la convergencia de los métodos de Jacobi y Gauss-Seidel.

Trivias

Trivia 1

P1: Consideremos una máquina con mantisa $m = 1$ y $-1 \leq l \leq 0$ en base 10. ¿Cuántos números tiene la máquina?

Solución: La máquina tiene 37 números, ya que si $m = 1$, entonces son de la forma $x = \pm 0.a_1 10^l$, claramente de 9 de la forma $0.a_1$, y 9 de la forma $0{,}0a_1$, entonces son $18 \cdot 2 + 1$ por los números negativos e incluyendo el 0.

P2: Consideremos una máquina con mantisa $m = 2$ y $0 \leq l \leq 2$. ¿Cuá es el número positivo más pequeño?

Solución: Por definición, son de la forma $x = 0.a_1 a_2 10^l$. Para toma el pequeño, claramente se toma $l = 0$, entonces sería 0,1 ya que estríctamente se pide que $a_1 \neq 0$, por lo que no puede ser 0,01.