Métodos Computacionales

PhD. Alejandro Segura

2021

Tabla de Contenido

Índice de figuras

Capítulo 1

Derivación e integración

1.1. Error de redondeo

Es la diferencia entre el valor exacto de un número y la aproximación calculada debida al redondeo. Por ejemplo, $\pi = 3{,}1415926535...$ si se aproxima a 3,1416 el error es $7{,}3464 \times 10^{-6}$, entonces la pregunta natural es: ¿cuál es el número más pequeño que podemos aproximar usando el computador? en otras palabras ¿cuál es el valor de ϵ para que se cumpla $1 + \epsilon = 1$.

1.2. Error de truncamiento

El error de truncamiento aparece cuando se usan aproximaciones en lugar de las expresiones exactas, en general, este tipo de error depende del tipo de aproximación que se realiza. Por ejemplo, cuando expandimos una cierta función alrededor de un punto y despreciamos términos de orden superior, se introducen error de truncamiento a nuestras estimaciones.

$$\begin{aligned} sin(x) &\cong x + \mathcal{O}(x^3) \\ sin(x) &\cong x - \frac{x^3}{3!} + \mathcal{O}(x^5) \end{aligned} \tag{1.1}$$

tiene diferente error de truncamiento para la estimación de la función $sin(x)$.

1.3. Derivación

Para construir la derivada numérica se define la siguiente discretización para nodos equi-espaciados.

$$x_j = x_0 + jh, \tag{1.2}$$

donde h se denomina paso, que es en general una variación "pequeña"de la función.

1.3.1. Derivada Progresiva

Dada esta condición podemos hacer una expansión en series de Taylor de $f(x)$ (el dominio son los puntos nodales).

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + ... \qquad \text{para} \qquad h << 1 \tag{1.3}$$

despejando la primera derivada tenemos:

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \underbrace{\frac{h}{2}f''(x)}_{\mathcal{O}(h)} \tag{1.4}$$

para algún punto de la partición:

$$f'(x_j) \cong \frac{f(x_{j+1}) - f(x_j)}{h} \tag{1.5}$$

La última expresión se denomina la derivada progresiva del punto x_j, la cuál tiene orden $\mathcal{O}(h)$ en la estimación.

1.3.2. Derivada Regresiva

Para obtener la derivada regresiva se expande:

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) + ... \qquad \text{para} \qquad h << 1 \tag{1.6}$$

despejando la primera derivada tenemos:

$$f'(x) = \frac{f(x) - f(x-h)}{h} + \underbrace{\frac{h}{2}f''(x)}_{\mathcal{O}(h)} \tag{1.7}$$

para algún punto de la partición:

$$f'(x_j) \cong \frac{f(x_j) - f(x_{j-1})}{h} \tag{1.8}$$

Para la derivada regresiva se tiene un orden de aproximación de orden $\mathcal{O}(h)$.

1.3.3. Derivada Central

Para estimar la derivada central se compara las expresiones de ambos desarrollos de Taylor.

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{3!}f'''(x) + ... \\ f(x-h) &= f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{3!}f'''(x) + ... \end{aligned} \tag{1.9}$$

Restamos las dos expresiones tenemos:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \underbrace{\frac{h^2}{3} f'''(x)}_{\mathcal{O}(h^2)} \tag{1.10}$$

para algún punto de la partición:

$$f'(x_j) \cong \frac{f(x_{j+1}) - f(x_{j-1})}{2h}. \tag{1.11}$$

Notar que la estimación central tiene un orden $\mathcal{O}(h^2)$ en la estimación.

1.3.4. Error Local

Una medida del error local es la distancia entre el valor estimado y el valor real.

$$\Delta_l(Df(x_j)) = f'(x_j) - \delta f_0(x_j) \tag{1.12}$$

1.3.5. Error global

Se define el error global como la propagación de errores locales en todos los puntos de la discretización.

$$\Delta_g(Df(x_j)) = \sqrt{\frac{\sum_{j=1}^{n} (f'(x_j) - \delta f_0(x_j))^2}{\sum_{j=1}^{n} f'(x_j)^2}} \tag{1.13}$$

1.3.6. Segunda Derivada

Para estimar la segunda derivada numérica, se suma los dos desarrollos de Taylor en la Ecuación (1.9).

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + \mathcal{O}(h^2) \tag{1.14}$$

para algún punto de la partición:

$$f''(x_j) \cong \frac{f(x_{j+1}) - 2f(x_j) + f(x_{j-1})}{h^2} \tag{1.15}$$

Notar que la estimación tiene un orden $\mathcal{O}(h^2)$ en la estimación.

1.3.7. Ejercicios

1. Es posible construir una aproximación de orden $\mathcal{O}(h^2)$ para la derivada progresiva y regresiva. Para tal propósito, escribir el polinomio de interpolación de grado 2, con x_1, x_2, x_3, siendo $y_j = f(x_j)$ (ver sección de interpolación de Lagrange). Usar el polinomio interpolador para mostrar que la derivada progresiva de orden dos es:

$$f'(x_1) \approx p'(x_1) = \frac{1}{2h}(-3f(x_1) + 4f(x_2) - f(x_3)) \tag{1.16}$$

más generalmente se puede escribir como:

$$f'(x) \cong \frac{1}{2h}(-3f(x) + 4f(x+h) - f(x+2h)) \tag{1.17}$$

Para $f(x) = \sqrt{tan(x)}$, estimar $f'(x)$ en el intervalo $[1, 2]$ con $h = 0{,}05$.

Hint: La derivada del polinomio interpolador es:

$$p'(x) = \frac{y_2 - y_1}{h} + \frac{1}{2h^2}(y_1 - 2y_2 + y_3)((x - x_1) + (x - x_2)) \tag{1.18}$$

2. Encuentre el operador $D^4 f(x_j)$. ¿Cuál es orden de la aproximación?

1.4. Método de Newton-Raphson

Es un método iterativo para encontrar las raíces reales polinomios usando conceptos de cálculo diferencial. Tomemos un punto cualquiera x_n y construimos la ecuación de la recta usando la derivada de $f(x)$ en x_n.

$$Df(x_n) = \frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n} \tag{1.19}$$

Se pretende que el siguiente punto en la iteración sea un raíz de $f(x_{n+1}) = 0$, por tanto:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{1.20}$$

Este es conocida como la forma iterativa de Newton-Raphson. Otro camino para deducir esta formula consiste en expandir $f(x)$ alrededor de x_n.

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{(x - x_n)^2}{2!} f''(x_n) + ... \tag{1.21}$$

Si se trunca la función hasta orden $\mathcal{O}(x^2)$ y se evalúa en el siguiente punto x_{n+1}, el cuál se considera un raíz de $f(x)$; se llega a la formula deseada.

1.4.1. Criterio de parada

Podemos usar el error relativo en cada iteración para detener el método.

$$\epsilon = \frac{|x_{k+1} - x_k|}{|x_{k+1}|} \tag{1.22}$$

el cual detiene el método cuando sea menor a una tolerancia, i.e, $\epsilon < 10^{-6}$.

1.4.2. Ejercicios

1. ¿De qué tipo es el error asociado a la estimación de raíces usando el método de Newton-Raphson?

2. ¿Cómo ajustar la precisión para estimar raíces con este método?

1.5. Método de bisección

Este método consiste en obtener una estimación de la raíces de una función, la cuales se suponen existe en cierto intervalo [a,b]. La existencia de las raíces (el teorema no establece cuántas raíces existen en dicho intervalo) se garantizan por medio del teorema de Bolzano, que establece lo siguiente:

Sea $f : [a, b] \to \mathbb{R}$ una función continua en [a,b] tal que $f(a) < 0 < f(b)$. Entonces, existe al menos un punto $c \in [a, b]$ tal que $f(c) = 0$.

Si se cumple el teorema de Bolzano en nuestro intervalo de estudio, podemos acercarnos a la raíz de la solución dividiendo el intervalo en su punto medio iterativamente.

$$x_m = \frac{a+b}{2} \tag{1.23}$$

De los dos intervalos elegimos el intervalo que cumpla la condición de Bolzano: $f(a)f(x_m) < 0$ o $f(x_m)f(b) < 0$. Repitiendo el proceso hasta que el intervalo sea lo suficientemente pequeño para acotar la raíz solución dentro de una vecindad arbitrariamente pequeña. La noción de distancia está dada por:

$$|x^k - x^{k-1}| < \epsilon^k \tag{1.24}$$

El algoritmo de bisección se puede describir como sigue:

a) En el intervalo [a,b] se debe cumplir la condición de Boltzano: $f(a)f(b) < 0$. De lo contrario, no existe la raíz en dicho intervalo.

b) De Divide el intervalo en su punto medio, donde tenemos dos intervalos: $[a, x_m]$ y $[x_m, b]$.

c) Si $f(x_m) = 0$, el punto medio es la raíz, con un cierto error y la iteración termina.

d) Si $f(a)f(x_m) < 0$. la raíz está en $[a, x_m]$. Hacer $b = x_m$ y repetir paso b). De lo contrario, la raíz está en $[x_m, b]$. Hacer $a = x_m$ y repetir el paso b).

Dado que en cada paso la longitud del intervalo se divide por dos, la longitud del intervalo después de k pasos se reduce a $\frac{b-a}{2^k}$. La aproximación a la raíz será el punto medio de dicho intervalo, de este modo, el error en la aproximación está dada por:

$$|\epsilon^k| < \frac{1}{2}\frac{b-a}{2^k}. \tag{1.25}$$

$|\epsilon^k|$ es asignado arbitrariamente. Para cierta precisión se tiene la siguiente expresión del número de pasos que acotan la solución en esa vecindad ($|x^k - x_{exacta}| < \epsilon^k$):

$$k > \frac{ln((b-a)/\epsilon^k)}{ln(2)} - 1 \tag{1.26}$$

1.6. Interpolación de Lagrange

Descubierto por Edwarg Waring en 1779 y redescubierto por Leonhard Euler en 1783, fue publicado por Lagrange en 1795. Se plantea como sigue: dado un conjunto de $n+1$ puntos diferentes $\Omega = \{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$, existe un polinomio interpolador de grado n:

$$p_n(x) = \sum_{i=0}^{n} f(x_i)\mathcal{L}_i(x), \tag{1.27}$$

donde $\mathcal{L}_i(x)$ es la base de Lagrange (también conocidas como funciones cardinales).

$$\mathcal{L}_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{1.28}$$

Este polinomio cumple que $p(x_k) = y_k$ para todo k en $\{0, ..., n\}$.

Ejemplo:
Encontrar las funciones cardinales ($\mathcal{L}_i(x)$) y el polinomio interpolador para el siguiente conjunto: $\Omega = \{(5, 10), (10, 15)\}$.

$$\mathcal{L}_0(x) = \frac{x - 10}{5 - 10} = -\frac{1}{5}(x - 10) \tag{1.29}$$

$$\mathcal{L}_1(x) = \frac{x - 5}{10 - 5} = \frac{1}{5}(x - 5) \tag{1.30}$$

Por tanto, el polinomio interpolador es:

$$\begin{aligned} p_1(x) &= 10\mathcal{L}_0(x) + 15\mathcal{L}_1(x) \\ p_1(x) &= x + 5 \end{aligned} \tag{1.31}$$

1.6.1. Error

Sea $f : I \to \mathbb{R}$, $\{x_i\}_{i=0}^n \subseteq I$, $x_i \neq x_j$ para $i \neq j$ y suponemos que f es derivable $n + 1$ veces. El error asociado a la interpolación está dado por:

$$E = f(x) - p(x) = \frac{f^{n+1}(\xi_x)}{(n + 1)!}(x - x_0)(x - x_1)...(x - x_n) \tag{1.32}$$

donde $p(x)$ es el polinomio interpolador en $\{x_i\}_{i=0}^n$ y $\xi_x \in$ al intervalo que contiene los puntos.

proof:
Si x es un punto x_k la identidad se satisface para cualquier ξ. De lo contrario, si x es fijo y diferente x_k se considera una función auxiliar:

$$F(t) = f(t) - p(t) - cL(t), \qquad \text{donde} \qquad c = \frac{f(x) - p(x)}{L(x)} \tag{1.33}$$

$L(x) = \prod_{i=0}^{n}(x - x_i)$. Si evaluamos la función auxiliar en los puntos x_k, $F(x_k) = y_k - y_k - 0 = 0$ para todo k. Por tanto, F tiene n+1 ceros. Adicionalmente $F(x) = f(x) - p(x) - cL(x) = 0$, dada

la definición de c. entonces la función F tiene $n+2$ ceros en el intervalo I. Ahora, por el teorema de Rolle, F' debe tener al menos n+1 ceros en el intervalo que tiene a los puntos x_k; entonces la (n+1)-ésima derivada debe tener al menos un cero. Sea ξ_x ese cero. Entonces derivamos $(n+1)$ veces y evaluamos en ξ_x:

$$F^{n+1}(\xi_x) = f^{n+1}(\xi_x) - c(n+1)! = 0 \tag{1.34}$$

debido a que la (n+1)-ésima derivada de $p(x)$ es cero. Entonces:

$$c = \frac{f^{n+1}(\xi_x)}{(n+1)!} \rightarrow cL(x) = f(x) - p(x) = \frac{1}{(n+1)!} f^{n+1}(\xi_x) L(x) \quad \blacksquare \tag{1.35}$$

1.6.2. Ejercicios

1. Demuestre que el polinomio interpolador es único.

2. Compruebe que las funciones cardinales son base (i.e, $L_i(x) = \delta_{ij}$ para cada $j \in \{0, 1, ..., n\}$).

3. ¿Con qué grado de exactitud podemos calcular $\sqrt{114}$ mediante la interpolación de de Lagrange para la función $f(x) = \sqrt{x}$, si elegimos los puntos $x_0 = 100$, $x_1 = 121$, $x_2 = 144$. Rpta: $|E| \simeq 1{,}8 \times 10^{-3}$.

4. En el lanzamiento de una bala, una cámara fotográfica registra las siguientes posiciones en metros respecto al arma homicida (tome $\vec{g} = -9{,}8\ m/s^2\ \hat{j}$):

x[m]	y[m]
1.4000	0.4007
3.5000	0.5941
5.6000	0.2980

Cuadro 1.1: Posiciones de la bala a lo largo de su trayectoria.

Estime el vector velocidad inicial, que estaría definido por la magnitud y dirección. Rpta: $V_0 = 10\ m/s$ y $\theta = 20°$. *Hint:* Encuentre el termino lineal y cuadrático de la interpolación y compare con la ecuación de trayectoria de la bala.

1.7. Integración

Para el calculo de integrales definidas existe una familia de métodos denominados *Métodos de Newton-Côtes*, los cuales se basan en cambiar el integrando $f(x)$ a un polinomio interpolador que aproxima a $f(x)$ en el intervalo de integración. El grado del polinomio interpolador queda definido por el número de puntos a considerar, por ejemplo, en los casos más simples de interpolación líneal (Regla de trapecio) e interpolación cuadrática (Regla de simpson), se tiene expresiones sencillas que son fáciles de implementar computacionalmente.

1.7.1. Método de trapecio simple

Para la integral definida:

$$I = \int_a^b f(x)dx, \tag{1.36}$$

el método de trapecios simple cambia el integrando por un polinomio interpolador de grado uno. Este polinomio interpolador esta definido en el conjunto $\Omega = \{(a, f(a)), (b, f(b))\}$ que finalmente conduce a la siguiente aproximación:

$$f(x) \approx p_1(x) = \frac{x-b}{a-b}f(a) + \frac{x-a}{b-a}f(b), \qquad \forall x \in [a, b]. \tag{1.37}$$

De modo que la integral tiene un valor aproximado:

$$I = \int_a^b f(x)dx \cong \int_a^b p_1(x)dx = \frac{b-a}{2}(f(a) + f(b)) \tag{1.38}$$

El error en la estimación está asociado al procedimiento de interpolación. Suponiendo que $f(x)$ es continua y derivable de clase C^2 en el intervalo [a,b]:

$$f(x) = p_1(x) + \epsilon(x), \tag{1.39}$$

donde

$$\epsilon(x) = \frac{f''(\xi)}{2}(x-a)(x-b), \qquad a \leq \xi \leq b. \tag{1.40}$$

Entonces el error asociado a la integración por le método del trapecio está dado por ($h = b-a$):

$$E = \int_a^b \epsilon(x)dx = -\frac{h^3}{12}f''(\xi) \tag{1.41}$$

De esta forma el error alcanza un valor máximo para algún valor de ξ, donde la segunda derivada de $f(x)$ se maximice; de modo que podemos acotar el error máximo en esta estimación.

$$|E| \leq \frac{h^3}{12} \underbrace{max}_{a \leq \xi \leq b} |f''(\xi)|. \tag{1.42}$$

Notar que el error es de orden $\mathcal{O}(h^3)$.

1.7.2. Método de trapecio compuesto

Para la integral definida:

$$I = \int_a^b f(x)dx, \tag{1.43}$$

el método de trapecios compuesto genera una partición *equi-espaciada* tal que: $x_{i+1} - x_i = h$, $\forall_i = [1, ..., n]$ sobre el conjunto $\Omega = \{(x_0, f(x_0)), (x_1, f(x_1)), ..., (x_n, f(x_n))\}$. Las condiciones de borde corresponden con los límites de integración ($x_0 = a$ y $x_n = b$), definiendo el paso de integración $h = \frac{b-a}{n}$. La integral se puede escribir como:

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + ... + \int_{x_{n-1}}^{x_n} f(x)dx, \tag{1.44}$$

aplicando el método de trapecios simple se tiene:

$$\int_a^b f(x)dx \approx \frac{h}{2}(f(x_0) + f(x_1)) + \frac{h}{2}(f(x_1) + f(x_2)) + ... + \frac{h}{2}(f(x_{n-1}) + f(x_n)), \tag{1.45}$$

tenemos la expresión para la regla de trapecio compuesta:

$$\int_a^b f(x)dx \approx \frac{h}{2}(f(a) + f(b)) + h\sum_{i=1}^{n-1} f(x_i). \tag{1.46}$$

La estimación del error se calcula sumando los errores en cada sub-intervalo.

$$E = \sum_{i=1}^{n} E_i = -\frac{h^2}{12}(f''(\xi_1) + f''(\xi_2) + ... + f''(\xi_n)) \tag{1.47}$$

El error puede ser acotado por el valor máximo promedio que tome la segunda derivada en el intervalo $[a, b]$.

$$|E| \leq \frac{h^2(b-a)}{12} \underbrace{max}_{a\leq\xi\leq b} |f''(\xi)|. \tag{1.48}$$

Notar que el error es de orden $\mathcal{O}(h^2)$.

1.7.3. Método de Simpson simple 1/3

Para la integral definida:

$$I = \int_a^b f(x)dx, \tag{1.49}$$

el método de Simpson simple cambia el integrando por un polinomio interpolador de grado dos. Este polinomio interpolador esta definido en el conjunto $\Omega = \{(a, f(a)), (x_m, f(x_m)), (b, f(b))\}$, donde $x_m = \frac{a+b}{2}$ es el punto medio del intervalo. La interpolación conduce a la siguiente aproximación:

$$f(x) \approx p_2(x) = \frac{(x-b)(x-x_m)}{(a-b)(a-x_m)}f(a) + \frac{(x-a)(x-b)}{(x_m-a)(x_m-b)}f(x_m) + \frac{(x-a)(x-x_m)}{(b-a)(b-x_m)}f(b), \quad \forall x \in [a, b]. \tag{1.50}$$

De modo que la integral tiene un valor aproximado:

$$\int_a^b f(x)dx \cong \int_a^b p_2(x)dx = \frac{h}{3}(f(a) + 4f(x_m) + f(b)). \tag{1.51}$$

Adicionalmente, es importante mencionar que *la discretización debe ser par*. El error de la aproximación tiene una característica interesante, dado que el error de la interpolación genera

un resultado idénticamente nulo. Suponiendo que $f(x)$ es continua y derivable de clase C^3 en el intervalo [a,b]:

$$f(x) = p_2(x) + \epsilon(x), \tag{1.52}$$

donde

$$E = \int_a^b \epsilon(x)dx = \int_a^b \frac{f'''(\xi)}{4!}(x-a)(x-b)(x-(a+b)/2)dx = 0, \qquad a \leq \xi \leq b. \tag{1.53}$$

Esto significa que la regla de Simpson es exacta a orden $\mathcal{O}(h^3)$, entonces necesitamos otro camino para calcular el error de la estimación a orden $\mathcal{O}(h^4)$. Para tal propósito, se va a expandir en series de Taylor alrededor del punto medio x_m a $f(x)$ y al polinomio interpolador $p_2(x)$. Suponiendo que $f(x)$ es continua y derivable de clase C^4 en el intervalo [a,b], a orden $\mathcal{O}(h^4)$ se tiene:

$$f(x) = f(x_m) + f'(x_m)(x-x_m) + \frac{f''(x_m)}{2!}(x-x_m)^2 + \frac{f'''(x_m)}{3!}(x-x_m)^3 + \epsilon(x), \tag{1.54}$$

donde el error es de orden $\mathcal{O}(h^4)$.

$$\epsilon(x) = \frac{f^{(4)}(\xi)}{4!}(x-x_m)^4. \tag{1.55}$$

Para reemplazar en la regla simple de Simpson, se debe encontrar $f(a)$ y f(b), entonces:

$$\begin{aligned} f(a) &= f(x_m-h) = f(x_m) + f'(x_m)(-h) + \frac{f''(x_m)}{2!}(-h)^2 + \frac{f'''(x_m)}{3!}(-h)^3 + \epsilon(x_m-h) \\ f(b) &= f(x_m+h) = f(x_m) + f'(x_m)(+h) + \frac{f''(x_m)}{2!}(+h)^2 + \frac{f'''(x_m)}{3!}(+h)^3 + \epsilon(x_m+h) \end{aligned} \tag{1.56}$$

Por tanto, la regla de Simpson queda expresada como (Notar como el término de orden $\mathcal{O}(h^3)$ no contribuye):

$$\begin{aligned} \frac{h}{3}(f(a)+4f(x_m)+f(b)) &= \frac{h}{3}(6f(x_m) + f''(x_m)h^2 + \epsilon(x_m+h) + \epsilon(x_m-h)) \\ &= \frac{h}{3}(6f(x_m) + f''(x_m)h^2 + \frac{1}{12}f^{(4)}(\xi)h^4) \end{aligned} \tag{1.57}$$

Ahora integrando el desarrollo de Taylor dado por la Ecuación (1.54):

$$\int_a^b f(x)dx = 2hf(x_m) + \frac{f''(x_n)}{3}h^3 + \frac{f^{(4)}(\xi)}{60}h^5. \tag{1.58}$$

De modo que es posible calcular el error como la diferencia en las dos integrales.

$$\begin{aligned} E &= \int_a^b f(x)dx - \int_a^b p_2(x)dx \\ &= \frac{f^{(4)}(\xi)}{60}h^5 - \frac{f^{(4)}(\xi)}{36}h^5 = -\frac{1}{90}f^{(4)}(\xi)h^5 \end{aligned} \tag{1.59}$$

El error puede ser acotado por el valor máximo promedio que tome la cuarta derivada en el intervalo $[a, b]$; notar que el error es de orden $\mathcal{O}(h^5)$.

$$|E| \leq \frac{1}{90}h^5 \underbrace{max}_{a\leq\xi\leq b} |f^{(4)}(\xi)|. \tag{1.60}$$

1.7.4. Método de Simpson compuesto

Para la integral definida:

$$I = \int_a^b f(x)dx, \tag{1.61}$$

el método de Simpson compuesto genera una partición *equi-espaciada* tal que: $x_{i+1} - x_i = h$, $\forall_i = [1, ..., n]$ sobre el conjunto $\Omega = \{(x_0, f(x_0)), (x_1, f(x_1)), ..., (x_n, f(x_n))\}$. Las condiciones de borde corresponden con los límites de integración ($x_0 = a$ y $x_n = b$). Definiendo el paso de integración como $h = \frac{b-a}{n}$, la integral se puede escribir:

$$\int_a^b f(x)dx = \int_a^{x_2} f(x)dx + \int_{x_2}^{x_4} f(x)dx + ... + \int_{x_{n-2}}^{b} f(x)dx \tag{1.62}$$

Notar que el número de puntos de la partición deber ser par y que los puntos medios serían el sub-conjunto $\{x_{2n+1}\}$. Aplicando el método simple a cada integral tenemos:

$$\int_a^b f(x)dx \approx \frac{h}{3}\bigg(f(a)+4f(x_1)+f(x_2)+f(x_2)+4f(x_3)+f(x_4)+...+f(x_{n-2})+4f(x_{n-1})+f(b)\bigg) \tag{1.63}$$

En general, podemos escribir la expresión final para el método de Simpson compuesto:

$$\int_a^b f(x)dx \approx \frac{h}{3}\bigg(f(a) + 4\sum_{i=1,impares}^{n-1} f(x_i) + 2\sum_{i=2,pares}^{n-2} f(x_i) + f(b)\bigg) \tag{1.64}$$

Para la estimación del error total, se debe considerar el error cometido en cada una de las integrales. De este modo:

$$|E| = \frac{1}{90}h^5\bigg(f^{(4)}(\xi_1) + f^{(4)}(\xi_2) + ... + f^{(4)}(\xi_{n/2}).\bigg) \tag{1.65}$$

Finalmente, acotando superiormente la estimación tenemos el error de la regla de Simpson compuesta a orden $\mathcal{O}(h^4)$.

$$|E| \leq \frac{b-a}{180}h^4 \underbrace{max}_{a\leq\xi\leq b} |f^{(4)}(\xi)|. \tag{1.66}$$

Cabe resaltar que la regla de Simpson compuesta que aproxima el integrando por un polinomio de orden tres, reduce en dos ordenes el error cometido por método de trapecio compuesto.

Ejemplo:
Para la integral:

$$\int_{-1}^{1} \sqrt{1+e^{-x^2}}dx, \tag{1.67}$$

Usando la regla de Simpson compuesta, ¿cuál debe ser el número de puntos para aproximar la integral con un error menor a 0.001?

Usando el operador discreto $D^4 f(x_j)$ se estima el máximo en $M = 3{,}183$. Entonces:

$$n \geq \sqrt[4]{\frac{(b-a)^5}{E180}M} \approx 4{,}87 = 6 \qquad \text{Intervalos} \tag{1.68}$$

debido que el número de intervalos debe ser par!

1.7.5. Cuadratura Gaussiana

La idea es aproximar la integral a una suma ponderada de la función $f(x)$, que use los valores más óptimos posibles para su determinación. En general, la formula de cuadratura está dada por:

$$G_M(f) = \sum_{i=1}^{n} w_i^{(n)} f(x_i^{(n)}) \cong \int_{-1}^{1} f(x)dx. \tag{1.69}$$

Donde $\{w_i^{(n)}\}$ son los pesos de ponderación y los $\{x_i^{(n)}\}$ son los puntos de Gauss, ambos cantidades dependen del número de puntos que se elijan. A diferencia de los métodos anteriores, queremos encontrar una regla que nos permita maximizar el grado de precisión en nuestras estimaciones. En general, para calcular los coeficientes se exige que la regla sea exacta para el polinomio de mayor grado que se tenga. Tenemos entonces:

$$\int_{-1}^{1} x^k dx = \sum_{i=1}^{n} w_i^{(n)} (x_i^{(n)})^k \qquad k = 0, 1, ..., N \tag{1.70}$$

donde N es tan grande como sea posible. Necesitamos $2n$ condiciones para integrar polinomios de grado $2n-1$, esto se denomina la regla de n puntos, para el grado polinomial $2n-1$.

Lema 1: Si $N = 2n$, entonces no existe el conjunto $\{w_i^{(n)}\}$ tal que se verifique la igualdad.

Supongamos que existe dicho polinomio, si definimos $L(x) = \prod_{j=1}^{n}(x - x_j^n)^2$ entonces $L(x) \geq 0$ y por tanto la integral $\int_{-1}^{1} L(x)dx > 0$. Sin embargo, $\sum w_i^{(n)} L(x_i^{(n)}) = 0$, entonces hay una contradicción.

Lema 2: Sea $\{w_i^{(n)}\}$ los pesos y $\{x_i^{(n)}\}$ los puntos de Gaus tal que se cumpla la Ecuación (1.70) para $k = 0, 1, 2, ..., N = 2n-1$, los pesos deben satisfacer:

$$w_i^{(n)} = \int_{-1}^{1} L_i^{(n)}(x)dx, \tag{1.71}$$

con

$$L_i^{(n)}(x) = \prod_{k=1,k\neq i}^{n} \frac{x - x_k^{(n)}}{x_i^{(n)} - x_k^{(n)}} \tag{1.72}$$

Las bases cardinales son de grado $\leq 2n-1$ y $L_i^{(n)}(x_j) = \delta_{ij}$, como la regla de cuadratura debe ser exacta para los polinomios de grado $2n-1$, entonces:

$$\int_{-1}^{1} L_i^{(n)}(x)dx = \sum_{j=1}^{n} w_j^{(n)} L_i^{(n)}(x_j) = \sum_{j=1}^{n} w_j^{(n)} \delta_{ij} = w_i^{(n)} \tag{1.73}$$

Ahora, tenemos un polinomio de grado n, $P_n(x) = \prod_{i=1}^{n}(x - x_i^{(n)})$ y supongamos que tenemos un polinomio cualquier $Q(x)$ de grado $\leq n-1$, de modo que la multiplicación tiene grado $\leq 2n-1$. Por tanto la identidad de la cuadratura se debe satisfacer.

$$\int_{-1}^{1} P_n(x)Q(x)dx = \sum_{i=1}^{n} w_i^{(n)} P_n(x_i^n) Q(x_i^{(n)}). \tag{1.74}$$

Por la definición de $P_n(x)$ la integral es nula y los polinomios deben formar una base ortonormal. Adicionalmente, el conjunto $\{x_i^n\}$ deben ser las raíces de los $P_n(x)$.

Ejemplo:

Obtener la regla de cuadratura para de $f(x)$ en el intervalo $[-3, 3]$ usando la siguiente partición $P = \{-1, 0, 1\}$.

Calculando las funciones de Lagrange:

$$\begin{aligned} L_1(x) &= \frac{1}{2}x(x-1) \\ L_2(x) &= -(x+1)(x-1) \\ L_3(x) &= \frac{1}{2}x(x+1) \end{aligned} \tag{1.75}$$

Integrando dichas funciones en el intervalo $[-3, 3]$ obtenemos el conjunto de pesos:

$$w_i^3 = \{9, -12, 9\} \tag{1.76}$$

Entonces la regla de cuadratura es:

$$\int_{-3}^{3} f(x)dx \approx 3[3f(-1) - 4f(0) + 3f(1)]. \tag{1.77}$$

Note que es exacta para $f(x) = x^2$.

1.7.6. Cuadratura de Gaus-Legendre

De la necesidad del uso de polinomios ortonormales y el cálculo de sus raíces, surge la cuadratura de Gaus-Legendre. Este calculo usa los polinomios de Legendre como conjunto ortonormal.

$$p_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n, \tag{1.78}$$

Los primeros polinomios son:

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= x \\ p_2(x) &= \frac{1}{2}(3x^2-1) \\ p_3(x) &= \frac{1}{2}(5x^3-3x) \end{aligned} \tag{1.79}$$

cumplen con la siguiente relación de completez:

$$\int_{-1}^{1} p_n(x)p_m(x)dx = \begin{cases} 0 & m \neq n \\ \frac{2}{2n+1} & m = n \end{cases} \tag{1.80}$$

y las siguientes relaciones de recurrencia.

$$(n+1)p_{n+1}(x) = (2n+1)xp_n(x) - np_{n-1}(x). \tag{1.81}$$

$$(x^2-1)\frac{dp_n}{dx} = nxp_n - np_{n-1}. \tag{1.82}$$

Adicionalmente se tiene el siguiente coeficiente director.

$$a_n = \frac{(2n)!}{2^n(n!)^2} \tag{1.83}$$

Para calcular los pesos de la cuadratura de Gauss-Legendre, se requiere escribir las funciones cardinales en términos de los polinomios de Legendre. Si definimos las funciones base de Lagrange como:

$$L(x) = \frac{p_n(x)}{x - x_k} \tag{1.84}$$

Tenemos una forma indeterminada cuando $x = x_k$. Para conocer dicho límite, si existe, usemos la regla de L'Hôpital.

$$\lim_{x\to x_i} L(x) = \left.\frac{\frac{dp_n(x)}{dx}}{\frac{d(x-x_k)}{dx}}\right|_{x=x_k} = p_n'(x_k) \tag{1.85}$$

como las funciones de Lagrange deben ser base y valor 1 justo en cada raíz, se tiene la siguiente expresión.

$$L(x) = \frac{1}{p_n'(x_k)}\frac{p_n(x)}{x - x_k} \tag{1.86}$$

Por tanto, los pesos pueden ser evaluados por:

$$w_k = \frac{1}{p_n'(x_k)} \int_{-1}^{1} \frac{p_n(x)}{x - x_k} dx \tag{1.87}$$

Ejemplo:

Calcular los pesos de ponderación para dos puntos ($n = 2$). El polinomio de Legendre sería $p_2(x) = \frac{1}{2}(3x^2 - 1)$, cuyas raíces son: $x_0 = \frac{1}{\sqrt{3}}$, $x_1 = -\frac{1}{\sqrt{3}}$. Los pesos están dados por:

$$w_0 = \frac{1}{3(\frac{1}{\sqrt{3}})} \int_{-1}^{1} \frac{\frac{1}{2}(3x^2 - 1)}{x - \frac{1}{\sqrt{3}}} dx = 1. \tag{1.88}$$

$$w_1 = \frac{1}{3(-\frac{1}{\sqrt{3}})} \int_{-1}^{1} \frac{\frac{1}{2}(3x^2 - 1)}{x + \frac{1}{\sqrt{3}}} dx = 1. \tag{1.89}$$

Por tanto, la regla de cuadratura para dos puntos es:

$$\int_{-1}^{1} f(x)dx = f(\frac{1}{\sqrt{3}}) + f(-\frac{1}{\sqrt{3}}) \tag{1.90}$$

para integrales que tiene un límite de integración diferente al de los polinomios de Legendre, es posible hacer un cambio de variable de modo que -1 coincida con el límite inferior y 1 con el límite superior de la integral. Planteando la siguiente proporción:

$$\frac{x - a}{b - a} = \frac{t - (-1)}{1 - (-1)}, \tag{1.91}$$

de manera que:

$$\begin{aligned} x &= \frac{1}{2}[t(b - a) + a + b] \\ dx &= \frac{1}{2}(b - a) \end{aligned} \tag{1.92}$$

Usando este cambio de variable, podemos escribir la regla de cuadratura para cualquier intervalo de integración.

$$\int_a^b f(x)dx = \frac{1}{2}(b - a) \sum_{k=0}^{n} w_k f(\frac{1}{2}[t_k(b - a) + a + b]) \tag{1.93}$$

La Ecuación (1.87) aunque es correcta, resulta poco útil para calcular los pesos a alto orden computacionalmente. Es posible encontrar una versión integrada de dicha expresión usando el siguiente teorema:

Theorem 1. *(Identidad de Christoffel-Darboux) Sea $\{p_{n(x)}\}_{n=0}^{\infty}$ el sistema ortonormal de polinomios respecto a un peso $\omega(x)$. Entonces, para todo $x, y \in R$, $x \neq y$ y $n \geq 1$ se cumple.*

$$\sum_{k=0}^{n-1} p_k(x)p_k(y) = \frac{a_{n-1}}{a_n} \left[\frac{p_n(x)p_{n-1}(y) - p_{n-1}(x)p_n(y)}{x - y} \right] \tag{1.94}$$

siendo $p_n(x) = a_n x^n + ...$ y $a_n > 0$

Usando esta identidad y haciendo $y = x_k$ las raíces de los polinomios tenemos.

$$\sum_{k=0}^{n-1} p_k(x)p_k(x_k) = \frac{a_{n-1}}{a_n}\frac{p_n(x)p_{n-1}(x_k)}{x - x_k} \tag{1.95}$$

Integrando, tenemos:

$$1 = \frac{a_{n-1}}{a_n}p_{n-1}(x_k)\int_{-1}^{1}\frac{p_n(x)}{x - x_k}dx \tag{1.96}$$

Por tanto, podemos escribir la Ecuación (1.87) como:

$$w_k = \frac{a_n}{a_{n-1}p_{n-1}(x_k)p'_n(x_k)} \tag{1.97}$$

Un cálculo sencillo muestra que $\frac{a_n}{a_{n-1}} = 2/n$ y usando la Ecuación (1.82).

$$w_k = \frac{2}{(1 - x_k^2)[p'_n(x_k)]^2}, \qquad k = 1, ..., n \tag{1.98}$$

Esta última expresión permite calcular los pesos usando librerias como `Simpy`.

1.7.7. Ejercicios

1. Hacer pasos intermedios para regla de trapecio simple, Ecuación (1.38).
2. Encontrar el error para regla de trapecio simple, Ecuación (1.41).
3. Hacer los pasos intermedios para encontrar la regla de Simpson simple, Ecuación (1.51).
4. Verificar el resultado presentado en la Ecuación (1.53).
5. La regla de Simpson 3/8 consiste en aproximar el integrando por un polinomio interpolador de orden 3. Use el paquete `Simpy` de `Python` para encontrar las funciones cardinales de dicha aproximación. Use los resultados para demostrar que:

$$\int_a^b f(x) \approx \frac{3h}{8}\left[f(a) + 3f(\frac{2a + b}{3}) + 3f(\frac{a + 2b}{3}) + f(b)\right]. \tag{1.99}$$

   Es importante mencionar que *la discretización debe ser múltiplo de tres.*

6. Muestre que error asociado a la regla de Simpson 3/8 simple está dado por:

$$E = \frac{f^{(4)}(\xi)}{4!}\int_a^b (x - x_0)(x - x_1)(x - x_2)(x - x_3)dx = -\frac{3}{80}h^5 f^{(4)}(\xi). \tag{1.100}$$

   Hint: Considere la siguiente integral:

$$\mathcal{I} = \int_0^{3h} (x)(x - h)(x - 2h)(x - 3h)dx \tag{1.101}$$

7. Muestre que para la regla de Simpson 3/8 compuesta, el error puede ser acotado por:

$$|E| \leq \frac{(b-a)}{80} h^4 \underbrace{max}_{a \leq \xi \leq b} |f^{(4)}(\xi)|. \tag{1.102}$$

8. Evaluar:

$$\int_1^2 \frac{1}{x^2} dx, \tag{1.103}$$

por medio de la cuadratura de Gauss-Legendre con dos y tres puntos. Rpta: $I_2 = 0{,}497041$, $I_3 = 0,499874$.

9. Escribir el polinomio $p(x) = 3 + 5x + x^2$ en la base de Legendre. Rpta: $p(x) = \frac{10}{3} p_0(x) + 5p_1(x) + \frac{2}{3} p_2(x)$

10. En el problema de la radiación de cuerpo negro aparece la siguiente integral.

$$\int_0^\infty \frac{x^3}{e^x - 1} dx = \frac{\pi^4}{15}. \tag{1.104}$$

a) Calcular la integral usando el método de la cuadratura Gauss-Laguerre para $n = 3$ puntos.

b) Para esta estimación, gráfique el error relativo ($\epsilon_r(n) = I_{estimated}(n)/I_{exact}$) como una función del número de puntos, con $n = [2, 3, ..., 10]$.

Este método es adecuado para estimar integrales del tipo:

$$\int_0^\infty e^{-x} f(x) \approx \sum_{i=1}^{n} w_k f(x_k) \tag{1.105}$$

Hint: Los pesos para la cuadratura de Gauss-Laguerre están dados por:

$$w_k = \frac{x_k}{(n+1)^2 [L_{n+1}(x_k)]^2} \tag{1.106}$$

Capítulo 2

Método de MonteCarlo

2.1. Generación pseudo-aleatoria de números

En general, no es posible generar auténticos números aleatorios por computador. Para obtener una secuencia aparentemente aleatoria, se utilizan métodos deterministas con altos periodos de repetición. Los métodos de congruencia lineal están definidos por:

$$x_{n+1} = (ax_n + c) mod\ m \qquad n \geq 0, \tag{2.1}$$

donde $0 \leq a < m$ se denomina multiplicador, $0 \leq c < m$ es el incremento y m es el módulo del método. En la secuencia x_0 se denomina la semilla del generador y debe ser ajustada *inteligentemente* para obtener secuencias diferentes en cada generación. En la expresión congruencial, el modulo representa el resto entre $(ax_n + c)$ y m, Por ejemplo si $x_0 = 1$, $a = 16$, $c = 0$ y $m = 7$ tenemos la siguiente secuencia $x_1 = 2$, $x_2 = 4$, ... Uno de los generadores más conocidos y usados tanto en `C++` como en `Python` está definido por:

$$\begin{aligned} a &= \text{5DEECE66D (base 16)} \\ c &= \text{B (base 16)} \\ m &= 2^{48} \end{aligned} \tag{2.2}$$

Para obtener un secuencia de números normalizada entre cero y uno $(0 \leq r \leq 1)$, se debe normalizar al modulo (m) la ecuación congruencial.

$$r = \frac{x_{n+1}}{m} = \frac{(ax_n + c) mod\ m}{m} \tag{2.3}$$

Para obtener una distribución de número en otro intervalo cualquiera $(A \leq x \leq B)$ se tiene debe escalar la distribución $\mathcal{U}(0, 1)$.

$$x = A + (B - A)r. \tag{2.4}$$

Por otro lado, es importante garantizar la aleatoriedad de la secuencia, para tal propósito, se tienen algunas pruebas a la función de distribución de dicha secuencia [2], por ejemplo, podemos usar el momento de la distribución $(P(x))$:

$$\frac{1}{N}\sum_{i=1}^{N} x_i^k \cong \int_0^1 x^k P(x)dx \cong \frac{1}{k+1} + \mathcal{O}(\frac{1}{\sqrt{N}}) \tag{2.5}$$

También se puede estudiar las correlación entre los k-vecinos cercanos, donde k es pequeño:

$$C(k) = \frac{1}{N}\sum_{i=1}^{N} x_i x_{i+k}, \qquad k = 1, 2, ... \tag{2.6}$$

Si la secuencia es aproximadamente uniforme podemos aproximar la suma como:

$$\frac{1}{N}\sum_{i=1}^{N} x_i x_{i+k} \cong \int_0^1 dx \int_0^1 dy x y P(x,y) = \frac{1}{4}. \tag{2.7}$$

De modo que, si la distribución de números pseudo-aleatorio sigue una función de distribución uniforme, $C(k)$ debe tener como máximo $1/4$. La Figura [2.1] muestra las correlaciones de los primeros 30 vecinos para un generador de números aleatorios malo y uno bueno. Note que un buen generador de eventos debe mantener las correlaciones alrededor de $1/4$.

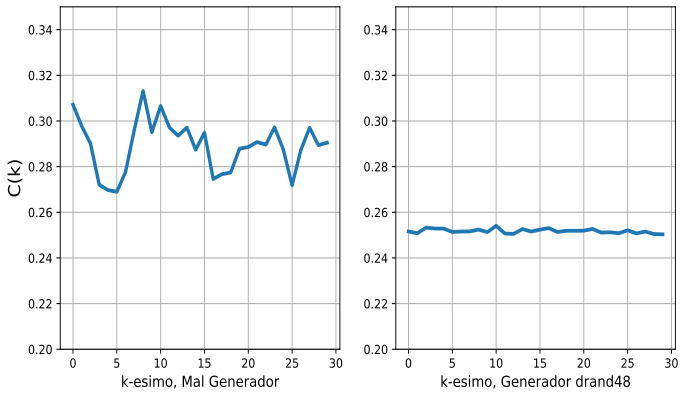

Figura 2.1: Correlaciones en los primeros k-vecinos para un mal generador (izquierda) y un buen generador (drand48) de números aleatorios.

2.2. Integración MonteCarlo

La integración Montecarlo es un método de integración de funciones generales que se sustenta en la ley de grandes números.

2.2.1. Ley de grandes números

Sea $\{x_i\}$ un conjunto de puntos muestrales (observaciones o datos simulados por computador) y el promedio muestral de dicho conjunto.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2.8}$$

Al aumentar el tamaño de la muestra, el promedio muestra converge al valor esperado con probabilidad i.

$$\begin{aligned} \text{Muestral} &\rightarrow \textit{Poblacional} \\ \bar{x} &\rightarrow \mathbb{E}(x) \end{aligned}$$

De esta manera, el operador integración $\int_0^1 f(x)dx$ puede verse como el valor esperado de $\mathrm{E}(f(x))$ donde $x \sim \mathcal{U}(0,1)$ tiene una distribución uniforme. De este modo, tomemos una muestra $x_1, x_2, ... \sim \mathcal{U}(0,1)$ y evaluamos en la función $f(x_1), f(x_2), ...$, entonces:

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \underbrace{\rightarrow}_{N\rightarrow\infty} \mathbb{E}(f(x)) \tag{2.9}$$

Otro resultado interesantes es que si x es una variable aleatoria con densidad de probabilidad f y g es una función $g : \mathbb{R} \rightarrow \mathbb{R}$, entonces:

$$\mathbb{E}(g(x)) = \int_{-\infty}^{\infty} g(x)f(x)dx = \sum_{i=1}^{N} g(x_i)f(x_i). \tag{2.10}$$

2.2.2. Método de la transformada inversa

Sea x una variable aleatoria con densidad de probabilidad f, y sea $F(x) = \int_{-\infty}^{x} f(s)ds$ la función acumulada de f. Dado que, la función acumulada de probabilidad es monotona creciente resulta ser inyectiva. De manera que, existe un único valor en el rango tal que:

$$F^{-1}(u) = x. \tag{2.11}$$

Por tanto, podemos generar una distribución uniforme y tomar la pre-imagen en la distribución acumulada. El resultado es el siguiente:

Sea $U \sim \mathcal{U}(0,1)$ y F una distribución acumulada, entonces F es injectiva en algún intervalo pre-imagen de [0,1]. Si X se define como $F^{-1}(U) = X$, entonces X tiene distribución F.

proof:

Queremos ver si $P(X \leq x) = F(x)$, lo que significa que x tiene distribución F. Por definición de X:

$$P(X \leq x) = P(F^{-1}(U) \leq x) \tag{2.12}$$

Como F es monotona creciente en el intervalo [0,1] es invertible:

$$\begin{aligned} P(X \leq x) &= P(F(F^{-1}(U)) \leq F(x)) \\ &= P(U \leq F(x)) \\ &= F(x) \end{aligned} \tag{2.13}$$

El último paso se justifica dado que U tiene distribución uniforme $U \sim \mathcal{U}(0,1)$. El algoritmo se resume como sigue:

1. Generar un número aleatorio que siga una distribución uniforme $U \sim \mathcal{U}(0,1)$.
2. Tomar la pre-imagen de U, $X = F^{-1}(U)$.

2.2.3. Método de aceptación y rechazo

En la gran mayoria de los casos es imposible encontrar la función inversa para usar el método de transformada inversa. En estos casos, se utiliza el método de aceptación y rechazo, el cual mediante la distribución uniforme de números puede calcular el área bajo la curva de distribuciones complicadas. Este método consiste se base en los siguientes pasos:

1. Generamos un x_i que siga una distribución uniforme entre contenido en el intervalo de integración [a,b].
2. Para x_i, generamos un y_i que siga una distribución uniforme entre 0 y el máximo de f(x).
3. Si $y_i < f(x_i)$ incluimos el valor x_i en la lista, de otro modo, no incluimos el valor x_i.

2.2.4. Incertidumbre del método de MonteCarlo

La incertidumbre en el método de MonteCarlo se estima propagando el error cometido en cada uno de los puntos muestrales x_i. Recordemos que la media muestral está dada por:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.14}$$

entonces:

$$\frac{\partial \bar{x}}{\partial x_i} = \frac{1}{N} \tag{2.15}$$

Teniendo en cuenta que cada medición es independiente la varianza asociada a la media muestra es:

$$Var(\bar{x}) = \sum_{i=1}^{N} \left(\frac{\partial \bar{x}}{\partial x_i} \right)^2 \delta_i^2 = \sum_{i=1}^{N} \frac{1}{N^2} \delta_i^2 \tag{2.16}$$

Por tanto, la incertidumbre está dada por:

$$\sqrt{Var(\bar{x})} \cong \frac{\delta}{\sqrt{N}} \tag{2.17}$$

δ es despreciable con el número de eventos simulados, en general se ajusta a 1 para tener una relación exacta con $1/\sqrt{N}$.

En esta ecuación es importante resaltar que la precisión del Método disminuye con $\sqrt{N}$, lo cuál debe tenerse en cuenta para cada calculo realizado con esta técnica. En el cálculo del número aureo usando la sucesión Fibonnaci aunque los puntos no provienen de un proceso aleatorio, la precisión del método tiene una dependencia similar.

2.2.5. Metropolis-Hasting algorithm

Suppose you wish to sample the random variable $x \sim f(x)$, but you can not use the direct simulation, the inverse CDF or accept-reject methods. But you can evaluate $f(x)$ at least up to a proportionality constant; then you can use the Metropolis-Hastings algorithm.

Let $f(x)$ be the (possible unnormalized) target density, x^j be a current value and $q(x/x^j)$ be a proposal distribution, which might depend on the current value x^j. The algorithm is summarized as follows:

a) Sample $x^* \sim q(x/x^j)$.

b) Calculate the acceptance probability:

$$\rho(x^j, x^*) = min\left[1, \frac{f(x^*)q(x^j/x^*)}{f(x^j)q(x^*/x^j)}\right], \tag{2.18}$$

if $\rho(x^j, x^*) > r$, x^* is accepted. Where $r \sim \mathcal{U}(0, 1)$.

c) Set $x^{j+1} = x^*$ with probability $\rho(x^j, x^*)$, otherwise set $x^{j+1} = x^j$.

In general, this sequence is not independent, moreover, we concentrate on symmetry chains (random walks) where $q(x/x^j) = q(x^j/x)$. The acceptance probability is, therefore:

$$\rho(x^j, x^*) = min\left[1, \frac{f(x^*)}{f(x^j)}\right] \tag{2.19}$$

How this algorithm works?

Let $D_n(x)$ be the points density x_n around x and $D_n(y)$ the points density y_n around y. x will be accepted or rejected. We can quantify the flux of points on the neighborhood x as:

$$\delta D_n(x) = \delta D_{n+1}(x) - \delta D_n(x) \tag{2.20}$$

$$\delta D_n(x) = \underbrace{\sum_y D_n(y)P(y \to x)}_{Gain} - \underbrace{\sum_y D_n(x)P(x \to y)}_{Lost} \tag{2.21}$$

$$\sum_y D_n(y)P(x \to y)\left[\frac{P(y \to x)}{P(x \to y)} - \frac{D_n(x)}{D_n(y)}\right] \tag{2.22}$$

We have two properties:

1. The equilibrium is achieved asymptotically.

2. There is flux toward y, which tends $D_n(x)$ to the equilibrium.

$$\frac{P(y \to x)}{P(x \to y)} < \frac{D_n(x)}{D_n(y)} \tag{2.23}$$

As a reminder $\rho(x, y) > r$ is the acceptance condition. We can calculate $P(x \to y)$ as $p_{xy}A_{xy}$, where p_{xy} is the probability of generation of y given x, and A_{xy} is the acceptance probability. Since the Metropolis estrategy is symmetric ($p_{xy} = p_{yx}$), we get:

$$\frac{P(y \to x)}{P(x \to y)} = \frac{A_{yx}}{A_{xy}} = \rho(x, y) = \frac{f(x)}{f(y)} \tag{2.24}$$

Therefore, the equilibrium distribution is given by:

$$D_{eq} = D(x) = f(x) \tag{2.25}$$

2.2.6. Estimación por máxima verosimilitud (Likelihood)

Sea $\mathcal{A}(x_1, x_2, ..., x_n)$ un vector aleatorio cuya dsitribución depende de un parámetro θ. Definimos la función de verosimilitud de $\mathcal{A}$ como:

$$\mathcal{L}(\theta) = f_{x_1,x_2,...x_n}(x_1, x_2, ...x_n; \theta) \tag{2.26}$$

que es la función de densidad de probabilidad evaluada en el vector $\mathcal{A}$ y es función del parámetro; note que no deben ser idénticamente distribuidas. En caso que sean variables idénticamente distribuidas tenemos:

$$\begin{aligned} \mathcal{L}(\theta) &= f(x_1; \theta) f(x_2; \theta) ... f(x_n; \theta) \\ &= \prod_{i=1}^{n} f(x_i; \theta). \end{aligned} \tag{2.27}$$

Este método de estimación de parámetros consiste en maximizar la función $\mathcal{L}(\theta)$, el valor del parámetro que maximiza la función se denota por $\hat{\theta}$ y se denota como estimador máximo verosímil.

Ejemplo:

Sea $\mathcal{A}(x_1, x_2, ..., x_n)$, donde $\mathcal{A} \sim e^{\alpha}$.

$$\mathcal{L}(\alpha) = \prod_{i=1}^{n} \frac{1}{\alpha} e^{-x_i/\alpha} = \left(\frac{1}{\alpha}\right)^n e^{-\sum_{i=1}^{n} x_i/\alpha} \tag{2.28}$$

Dado que la función logaritmo natural es monótona, podemos maximizar $Ln(\mathcal{L}(\alpha))$.

$$Ln(\mathcal{L}(\alpha)) = nLn(1/\alpha) - \sum_{i=1}^{n} \frac{x_i}{\alpha} \tag{2.29}$$

Encontrando el máximo de esta función, tenemos:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X} \tag{2.30}$$

Ejemplo:

Sea $\mathcal{A}(x_1, x_2, ..., x_n)$, donde $\mathcal{A} \sim Pois(\lambda)$.

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} \tag{2.31}$$

Tomando logaritmo natural tenemos:

$$Ln(\mathcal{L}(\lambda)) = -n\lambda + \sum_{i=1}^{n} x_i Ln(\lambda) - Ln(\prod_{i=1}^{n} x_i!) \tag{2.32}$$

El siguiente sería el parámetro máximo verosímil.

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X} \tag{2.33}$$

2.2.7. Varianza de los estimadores

Vamos a expandir la función de verosimilitud alrededor del estimador, lo que significa que estamos en el punto máximo.

$$Ln(\mathcal{L}(\theta)) \cong Ln(\mathcal{L}(\hat{\theta})) + \frac{\partial Ln(\mathcal{L}(\theta))}{\partial\theta}\bigg|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2}\frac{\partial^2 Ln(\mathcal{L}(\theta))}{\partial\theta^2}\bigg|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 \tag{2.34}$$

Dado que estamos en el máximo:

$$Ln(\mathcal{L}(\theta)) \cong Ln(\mathcal{L}(\hat{\theta})) - \frac{(\theta - \hat{\theta})^2}{2\hat{\sigma}^2} \tag{2.35}$$

donde hemos definido:

$$\hat{\sigma}^2 = -\left(\frac{\partial^2 Ln(\mathcal{L}(\theta))}{\partial\theta^2}\right)^{-1} \tag{2.36}$$

Note que si evaluamos la función de verosimilitud en $\theta = \hat{\theta} \pm \hat{\sigma}$ obtenemos

$$Ln(\mathcal{L}(\hat{\theta} \pm \hat{\sigma})) = Ln(\mathcal{L}(\hat{\theta})) - 1/2 \tag{2.37}$$

Esto significa que a 1σ de desviación estándar el logaritmo de la función de verosimilitud evaluada desde el máximo debe cambiar en 1/2.

2.3. Ejercicios

1. Demostrar la Ecuación (2.4).

2. Programar un mal generador de números aleatorios y usar drand48 en `Python` para reproducir la Figura [2.1].

3. La distribución Beta está dada por:

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \leq x \leq 1 \tag{2.38}$$

   donde $\Gamma(n) = (n-1)!$. Para f(x; 2,4), halle el área bajo la curva usando el método de aceptación y rechazo con una incertidumbre del 1 %.

4. Sea $\mathcal{A}(x_1, x_2, ..., x_n)$, donde $\mathcal{A} \sim \mathcal{N}(\mu, \sigma)$. Muestre que los estimadores máximo verosímiles son:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned} \tag{2.39}$$

5. Para el ejercicio anterior, muestre que la matriz Hessiana evaluada en los estimadores es:

$$H(\hat{\mu}, \hat{\sigma}^2) = \begin{pmatrix} -n(\hat{\sigma}^2)^{-1} & 0 \\ 0 & -\frac{n}{2}(\hat{\sigma}^2)^2 \end{pmatrix}$$

   ¿Es esta matriz definida positiva o negativa?

6. La siguiente integral multidimensional:

$$\int_0^1 \cdots \int_0^1 2^{-7} \left(\sum_{i=1}^{8} x_i \right)^2 dx_1 dx_2 ... dx_8, \tag{2.40}$$

   tiene el valor exacto $\frac{25}{192}$, usando el método de MonteCarlo estime esta integral con tres cifras de precisión.

7. Sampling of the Likelihood function using the Metropolis Hastings algorithm (Frequentist approach). Implemente el algoritmo de Metropolis Hastings para muestrear la distribución de likelihood como función de sigma ($\mathcal{L}(\sigma|data)$). Use esta distribución para estimar la desviación estándar en el intervalo de confianza del 68 % (i.e, σ_-^+).

   a) Descargue los datos usados en clase: MLikelihood Data.

   b) Escriba la función de Likelihood asociada a la distribución Gausiana.

   c) El código debe calcular el Likelihood asociado a los datos para cada σ en la cadena de Markov.

d) La condición de aceptación es $\alpha = \frac{\mathcal{L}(\sigma*|data)}{\mathcal{L}(\sigma_0|data)}$, donde $\sigma*$ corresponde al valor candidato y σ_0 al valor actual.

e) Se sugiere un número de grande de iteraciones $\sim 10^5$.

f) Dibuje el Histograma. ¿Cuál es la mediana y los errores asimétricos de esta estimación?

g) Compare con el valor del σ obtenido en la clase (which is obtained using a Bayesian approach).

8. En general la varianza de estimadores es no calculable:

$$V(\hat{\theta}) = E(\hat{\theta}^2) - E(\hat{\theta})^2 \tag{2.41}$$

En el caso de la distribución exponencial tenemos un valor analítico dado por:

$$\begin{aligned} V(\hat{\theta}) &= \int_0^\infty ... \int_0^\infty \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 \frac{1}{\theta} e^{-x_1/\theta} ... \frac{1}{\theta} e^{-x_n/\theta} dx_1 ... dx_n \\ &- \left[\int_0^\infty ... \int_0^\infty \left(\frac{1}{n}\sum_{i=1}^n x_i\right) \frac{1}{\theta} e^{-x_1/\theta} ... \frac{1}{\theta} e^{-x_n/\theta} dx_1 ... dx_n\right]^2 \\ &= \frac{\theta^2}{n} \end{aligned} \tag{2.42}$$

a) Intente encontrar este resultado analíticamente.

b) Con el método de MonteCarlo compruebe este resultado para un conjunto de n=20 variables aleatorias $x_i ... x_n \sim Exp(\theta = 2)$. Generar varias muestras de distribuciones exponenciales para tener un buen promedio en el ensamble, por ejemplo: $N = 10^6$ (Se obtiene algo como $Var(\hat{\theta}) = 0{,}199$).

9. En física de altas energías el siguiente *Toy Model* es relevante. Se tiene un número observado de eventos totales ($n = 10$), un número de eventos de física conocida $b = 9$ y un número de eventos esperados de nueva física $s = 4$. La distribución está caracterizada por la siguiente función de Likelihood:

$$\mathcal{L}(\mu, \epsilon) = Poisson(n; \mu s + \epsilon b) Gauss(\epsilon; 1., 0{,}1) \tag{2.43}$$

$$\mathcal{L}(\mu, \epsilon) = \frac{1}{n!} e^{-(\mu s + \epsilon b)} (\mu s + \epsilon b)^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\epsilon-1)^2}{2\sigma^2}} \tag{2.44}$$

Donde ϵ es la eficiencia de reconstrucción de los eventos de física conocida tiene valor medio $\bar{\epsilon} = 1.$ y desviación estándar $\sigma = 0{,}1$. La contribución Gaussiana modela el error sistemático asociado a la estimación de b. μ se conoce como *signal strength*, el cuál es muy importante para descartar o no, nuevas teorías es física. Con esta información:

a) Reproduzca las gráficas de la Figura [2.2]. Los parámetros pueden variar como $0. < \mu < 2.$ y $0. < \epsilon < 2.$

b) Usando el algoritmo de Metropolis-Hastings encuentre $\hat{\mu}$ y $\hat{\epsilon}$ y los errores asociados a cada parámetro.

c) Elimine el error sistemático dado por la parte gaussiana. ¿Cuál sería el valor de $\hat{\mu}$?. A qué conclusiones llega?

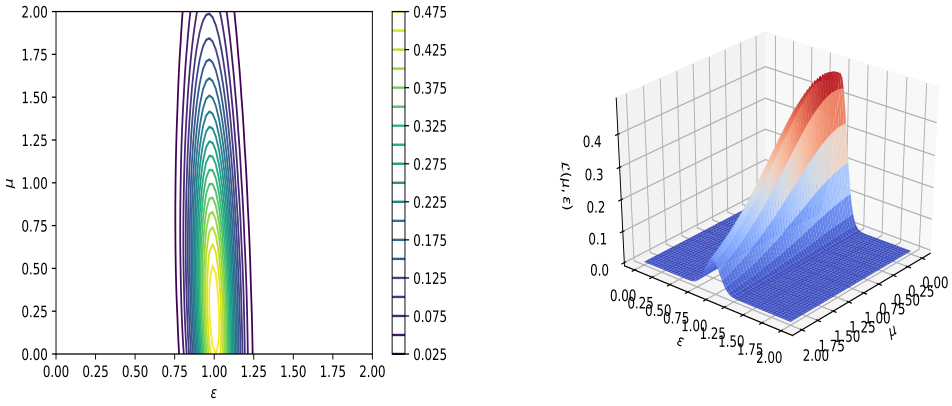

Figura 2.2: Gráfica de curvas de nivel de la función $\mathcal{L}(\mu, \epsilon)$ (izquierda) y superficie (derecha).

Capítulo 3

Statistics

La estadística estudia la organización y análisis de datos para dar conclusiones y tomar decisiones válidas. Es imposible estudiar el grupo completo de datos denominado población, en su lugar, se toma un sub-conjunto llamado muestra (que debe ser representativo). En general, la población puede ser finita o infinita. La descripción y análisis de unicamente la muestra, se denomina estadística descriptiva. Si por el contrario se desea inferir conclusiones sobre la población, con base a la muestra, se denomina estadística inferencial.

3.1. Distribución de frecuencias

Una de las operaciones más sencillas de realizar en un conjunto de datos es la ordenación en términos de la frecuencia de repetición de los datos. Para tal fin, vamos a definir algunos parámetros de la muestra que nos permite ordenar la información.

1. Rango de la población: Es la diferencia entre el mayor y el menor valor encontrado en la muestra. Por ejemplo, se mide el peso de 300 estudiantes donde el valor mayor es 100 kg y el menor es 65.4 kg. El rango de la muestra es: 34.6 kg.

2. Frecuencia de clase: Es apropiado definir categorías o clases (C_i) en los datos para separar la muestra en n clases. El número de datos pertenecientes a la clase C_i se denomina frecuencia de clase. Las clases tienen las siguientes propiedades:

   a) $C_i \neq \emptyset$
   b) $C_i \cap C_j = \emptyset$
   c) $\bigcup_{i=1}^{n} C_i = \Omega$

3. Límite de clase: En el ejemplo del peso de los estudiantes, si tomáramos la clase C_1 como todos los estudiantes que pesan entre 65 y 75 kg, entonces los límites de clase sería precisamente dichos valores.

4. Tamaño de clase: Es la diferencia entre los límites de clase, por ejemplo: 75 - 65 kg = 10 kg es el tamaño de clase de C_1.

5. Marca de clase: Es el punto medio de cada clase, el cuál se obtiene promediando el valor de los límites de clase.

Con estas definiciones podemos definir el concepto de histograma. Un histograma es el conjunto de clases, que pueden ser representadas gráficamente como los rectángulos, cuya base esta centrada en las marcas de clase y su altura es la frecuencia de clase. Note que, el área total histograma es proporcional al tamaño total de la muestra. *En ciertos casos resulta muy útil hacer que el área del histograma sea igual a 1 para comparar la forma de la distribución con otras distribuciones.*

Así mismo, podemos definir una distribución de frecuencias acumulada, que nos da información de como se acumula la información en los datos para valores menores a una marca de clase específica. La grafica de la distribución acumulada de frecuencia se denomina ojiva. Imaginemos la siguiente situación:

Peso [kg]	Frecuencia		Acumulada
60 - 70	6	Menor a 60	0
70 - 80	35	Menor a 70	6
80 - 90	40	Menor a 80	41
90 - 100	19	Menor a 90	81
		Menor a 100	100

Cuadro 3.1: Distribución de pesos y distribución acumulada de pesos para un conjunto de mediciones de 100 estudiantes.

Una descripción sencilla de la distribución acumulada es que 81 estudiantes tienen un peso menor a 90 kg.

3.2. Estadísticos

Sea X una variable aleatoria y sean $\Omega = \{X_1, X_2, ..., X_n\}$ variables aleatorias con la misma distribución que X. Diremos que el conjunto Ω es una muestra aleatoria de tamaño n de X. Dada una muestra de una población, un estadístico es una función real de la muestra:

$$T = f(X_1, X_2, ..., X_n) \tag{3.1}$$

como el estadístico es una variable aleatoria, entonces será en si mismo una variable aleatorio. Esto significa que tendrá una distribución, una media, etc. En general, la distribución que siga el estadística se denomina distribución muestral.

3.3. Pruebas de Hipótesis

El procedimiento estadístico para determinar si los datos de una muestra son compatibles con las características de una población. Para contextualizar el problema se realizan las siguientes definiciones:

- Una hipótesis estadística es una proposición sobre los parámetros de una población.
- Se denomina hipótesis nula a la hipótesis que se desea contrastar, la denominamos H_0. La hipótesis alternativa la denominamos con H_1.

- Un estadístico de prueba es un estadístico utilizado para determinar si se rechaza o no la hipótesis nula H_0.
- Región de no rechazo es el conjunto de valores del estadístico de prueba, para los cuales no rechazamos la hipótesis nula. Su región complementaria se llama la región de rechazo de H_0 en favor de H_1.
- Dado que hay una probabilidad de que el estadístico genere un valor en la región de rechazo, se tienen dos tipos de errores: Error tipo-1 es rechazar la hipótesis nula cuando es verdadera y el Error tipo-2 es no rechazar la hipótesis nula cuando es falsa.
- El nivel de significación del test determina la probabilidad de que el estadístico genere un valor en la región de rechazo de H_0. Se denomina α y está restringido al intervalo $0 \leq \alpha \leq 1$.

3.4. Ejercicios

3.4.1. Conteo

Realizar los siguientes cálculos estableciendo si es variación, permutación o combinación; con o sin repetición.

1. Carlos, Manuel, Sandra correrán los 100 metros planos. ¿De cuántas formas puede quedar el podio de primer y segundo lugar? Solo competirán ellos tres. R:6
2. ¿De cuántas formas se puede preparar una ensalada de frutas con solo 2 ingredientes, si se cuenta con plátano, manzana y uva? R:3
3. ¿De cuantas formas pueden hacer cola 5 amigos para entrar al cine? R:120
4. ¿De cuántas formas puede un juez otorgar el primero, segundo y tercer premio en un concurso que tiene 8 participantes R:336
5. El capitán de un barco solicita 2 marineros para realizar un trabajo, sin embargo, se presentan 10. ¿De cuántas formas podrá seleccionar a los 2 marineros? R:45
6. Eduardo tiene 7 Libros, ¿De cuántas maneras podrá acomodar cinco de ellos de un estante? R:2520
7. En un salón de 10 alumnos, ¿de cuántas maneras se puede formar un comité formado por 2 de ellos? R:45
8. ¿Cuántas palabras diferentes se puede formar con las letras de la palabra REMEMBER? R:1680
9. Un club de basketball tiene 12 jugadoras, una de ellas es la capitana María. ¿Cuántos equipos diferentes de 6 jugadoras se pueden formar, sabiendo que en todos ellos siempre debe estar la capitana María. R:462
10. Con 4 frutas diferentes, ¿cuántos jugos surtidos se pueden preparar?. Un jugo surtido se debe preparar con al menos 2 frutas. R:11

3.4.2. Probabilidad

Algunos ejercicios de probabilidad fueron del libro de probabilidad y estadística [3].

1. La calificación de matemáticas de 80 estudiantes están en el siguiente archivo: `https://github.com/asegura4488/DataBase/blob/main/MetodosComputacionalesReforma/Matematicas.txt`. Hallar la siguiente información:

   a) La calificación más alta.

   b) La más baja.

   c) El rango.

   d) La cinco más bajas.

   e) Las cinco más altas.

   f) La décima de mayor y menor.

   g) El número de estudiantes con calificación de 75 o más.

   h) El porcentaje de estudiantes con calificaciones mayores que 65 pero no superiores a 85.

   i) Las calificaciones de 0 a 100 que no aparecen en las notas.

2. La Tabla [3.2] muestra la distribución de frecuencia de los salarios de los empleados de una empresa anónima. Calcule la distribución acumulada de porcentajes y su respectiva gráfica.

Salarios	Número de empleados
250 - 259	8
260 - 269	10
270 - 279	16
280 - 299	14
290 - 300	10
300 - 309	5
310 - 319	2

Cuadro 3.2: Distribución de salarios para 65 empleados de la empresa anónima.

3. En el experimento de lanzar simultáneamente 3 dados de 6 caras calcule la probabilidad de obtener 1 par. En este caso programe 3 dados para hacer el experimento virtual con $n = 10^5$ veces. Escoja los casos favorables para calcular la probabilidad de dicho evento. $\lim_{n\to\infty} p(a) = \frac{5}{12}$

4. En el experimento de lanzar simultáneamente 5 dados de 6 caras calcule la probabilidad de obtener:

   a) 1 par.

   b) 2 pares distintos.
   En estos dos casos programe 5 dados para hacer el experimento virtual con $n = 10^5$ veces. Escoja los casos favorables para calcular la probabilidad de dicho evento. 1 par: $\lim_{n\to\infty} p(a) = \frac{25}{54}$ y 2 pares $\lim_{n\to\infty} p(b) = \frac{25}{108}$:

c) 4 de la misma cara: $p(b) = \frac{25}{1296}$

5. En un juego de Poker que consta de 5 cartas, encuentre la probabilidad de tener:

   a) 3 ases: $p(a) = \frac{94}{54145}$

   b) 4 cartas de corazones y 1 de bastos: $p(b) = \frac{143}{39984}$

6. Si se seleccionan al azar 3 libros de un estante que contiene 5 novelas. 3 libros de poemas y un diccionario, ¿cuál es la probabilidad de que:

   a) se tome el diccionario: $p(a) = \frac{1}{3}$

   b) se escojan 2 novelas y un libro de poemas: $p(b) = \frac{5}{14}$

7. En el experimento aleatorio del lanzamiento de 4 monedas determine la probabilidad de obtener dos sellos y dos caras, que llamaremos eventos A: $p(A) = \frac{3}{8}$. Realice el experimento computacional con $n = 10^5$ lanzamientos, etiquetando la opción cara con 1 y la opción sello con -1.

8. En el ejercicio anterior, imagine que las monedas están truncadas de tal manera que la probabilidad de que la moneda 1 sea cara es p_1 y que sea sello es $1 - p_1$. Usando el árbol de probabilidad, ¿cuál es la expresión de la probabilidad de obtener dos caras y dos sellos de este evento? Si el truncamiento de las monedas 1 y 2 puede variar como: $0{,}1 < p_1 < 0{,}9$ y $0{,}1 < p_2 < 0{,}5$, use el árbol de probabilidad para graficar la superficie de probabilidad del evento A. ¿En qué punto la probabilidad es mínima y máxima y cuáles son sus valores?

Capítulo 4

Simulación de n-cuerpos

4.1. Sistema de unidades.

Para empezar a hablar de simulación de n-cuerpos es importante primero definir el sistema de unidades que se van a usar para realizar dichas simulaciones. En el caso gravitacional, se hace pertinente usar el sistema Gaussiano de unidades. Recordemos el valor y unidades en el S.I. de la constante de gravitación universal:

$$G = 6{,}67 \times 10^{-11}, \qquad [G] = \frac{L^3}{T^2 M} \tag{4.1}$$

El sistema Gaussiano se basa en pasar las unidades fundamentales es términos de valores Astronómicos de nuestro sistema solar. La masa debe ser transformada a masa solares ($\mathrm{M}_\odot$), de modo que la masa del sol tiene valor 1. La distancia deber ser cambiada a la distancia media entre la tierra y el sol (au), de modo que la distancia entre el sol y la tierra vale 1. Finalmente, el tiempo debe ser transformado a días o años terrestres según convenga. Las cantidades son las siguientes:

$$\begin{aligned} \mathrm{M}_\odot &= 1{,}989 \times 10^{30}\ kg \\ au &= 1{,}496 \times 10^{11}\ m \\ dia &= 86400\ s \end{aligned} \tag{4.2}$$

Entonces las unidades de G serían.

$$[G] = 6{,}674 \times 10^{-11} \frac{m^3}{kgs^2} \times \left(\frac{1{,}989 \times 10^{30}\ kg}{1\mathrm{M}_\odot} \right) \times \left(\frac{86400s}{1dia} \right)^2 \left(\frac{1au}{1{,}496 \times 10^{11}\ m} \right)^3 \tag{4.3}$$

$$[G] = 2{,}9597 \times 10^{-4} \frac{au^3}{dias^2 \mathrm{M}_\odot} \tag{4.4}$$

Esta constante reduce enormemente el error de redondeo en el cálculo de sistemas gravitacionales. Adicionalmente, es posible encontrar un valor más intuitivo. Si recordamos la tercer ley de Kepler. Tenemos:

$$[G] = \frac{4\pi^2}{(1 \text{ año})^2 \mathrm{M}_\odot}(1au)^3 = \frac{4\pi^2}{(365{,}2421)^2}\frac{au^3}{dias^2 \mathrm{M}_\odot} \tag{4.5}$$

De modo que si usamos días terrestres usamos la Ecuación (4.4) y si usamos años terrestres como el tiempo corriendo en el computador, usamos $G = 4\pi^2$ que es un valor muy natural y fácil de recordar y programar. *Para el problema de la anomalía de Mercurio es necesario usar el tiempo en años terrestres para encontrar la velocidad de precesión adecuadamente.*

4.2. Sistemas Lineales

4.2.1. Método de Jacobi

El método de Jacobi es usado para resolver sistemas lineales del tipo $\mathbb{A}x = b$, el cuál construye una sucesión descomponiendo la matriz $\mathbb{A}$.

$$\mathbb{A} = \mathbb{D} + \mathbb{R} \tag{4.6}$$

donde $\mathbb{D}$ es la matriz diagonal y $\mathbb{R}$ es la suma de la matriz triangular inferior y la matriz triangular superior.

$$\begin{aligned} \mathbb{D}x &= \mathbb{R}x + b \\ x &= \mathbb{D}^{-1}(b - \mathbb{R}x) \end{aligned} \tag{4.7}$$

si $a_{ii} \neq 0$ tenemos un método iterativo.

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{i \neq j} a_{ij}x_j^k\right) \tag{4.8}$$

A nivel iterativo se necesitan todos los valores del vector x en el paso k para calcular el valor del vector en $k+1$.

Ejemplo: Sea el sistema lineal:

$$\begin{aligned} 3x - y - z &= 1 \\ -x + 3y + z &= 3 \\ 2x + y + 4z &= 7 \end{aligned} \tag{4.9}$$

encontrar la ecuación iterativa del método de Jacobi.

$$\mathbb{D}^{-1}b = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}\begin{pmatrix} 1 \\ 3 \\ 7 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1 \\ 7/4 \end{pmatrix} \tag{4.10}$$

$$\mathbb{D}^{-1}\mathbb{R} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}\begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1/3 & -1/3 \\ -1/3 & 0 & 1/3 \\ 1/2 & 1/4 & 0 \end{pmatrix} \tag{4.11}$$

Finalmente:

$$\begin{pmatrix} x^{k+1} \\ y^{k+1} \\ z^{k+1} \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1 \\ 7/4 \end{pmatrix} - \begin{pmatrix} 0 & -1/3 & -1/3 \\ -1/3 & 0 & 1/3 \\ 1/2 & 1/4 & 0 \end{pmatrix} \begin{pmatrix} x^{k} \\ y^{k} \\ z^{k} \end{pmatrix} \tag{4.12}$$

Este método se puede interpretar como una sucesión de vectores que ocurre en un punto arbitrario del espacio vectorial al vector solución, la cuál es generada por un operador de traslación $\mathbb{T}$ definido por el sistema de ecuaciones.

$$x^{k+1} = c + \mathbb{T}x^{k} \tag{4.13}$$

A nivel computacional para detener el método se calcula una métrica entre el vector solución y el vector calculado. Se define el residuo como:

$$Re = ||x - y|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4.14}$$

para detener el método se usa la siguiente condición:

$$||\mathbf{b} - \mathbb{A}\mathbf{x}|| < \delta \tag{4.15}$$

Convergencia: El método de Jacobi converge a la solución del problema lineal si la matriz asociada al sistema de ecuaciones es diagonal dominante.

Convergencia Fuerte: La sucesión $x^{k+1} = Tx^{k} + c$, para $k \geq 0$ converge a la sucesión única $x = Tx + c$ si y sólo si $\rho(T) < 1$. Definición de radio espectral:

Si $\lambda_1, ..., \lambda_n$ son los valores propios de una matriz A, entonces su radio espectral $\rho(A)$ se define como:

$$\rho(A) := \underbrace{max}_{i}(|\lambda_i|) \tag{4.16}$$

4.2.2. Método de Gauss-Seidel

EL método de Gauss-Seidel se basa en el método de Jacobi con la única diferencia que cada variable se actualiza con cada iteración interna. Usando el ejemplo anterior se evidencia como funciona:

$$\begin{aligned} x &= \frac{1 + y + z}{3} \\ y &= \frac{3 + x - z}{3} \\ z &= \frac{7 - 2x - y}{4} \end{aligned} \tag{4.17}$$

Si la iteración empieza en el vector nulo tenemos que los valores en la siguiente iteración son:

$$\begin{aligned} x &= \frac{1+0+0}{3} = 1/3 \\ y &= \frac{3+1/3+0}{3} = 10/9 \\ z &= \frac{7-2/3-10/9}{4} = 47/36 \end{aligned} \tag{4.18}$$

Teniendo la misma condición de convergencia del método de Jacobi, el método de Gauss-Seidel acelera el tiempo de convergencia a la solución. En general se tiene:

$$\begin{aligned} x^{k+1} &= \frac{1+y^k+z^k}{3} \\ y^{k+1} &= \frac{3+x^{k+1}-z^k}{3} \\ z^{k+1} &= \frac{7-2x^{k+1}-y^{k+1}}{4} \end{aligned} \tag{4.19}$$

4.3. Ejercicios

1. Escriba la velocidad de la luz $c = 3 \times 10^8\ m/s$ en unidades de au/año.

2. Usando el código visto en clase comprobar la tercera ley de Kepler para todos los planetas del sistema solar (incluyendo a Plutón). Tomar los semi-ejes mayores y excentricidad de Internet. Con el periodo de los planetas hacer un ajuste lineal entre T^2 y a^3. ¿Cuál es el valor de la constante de proporcionalidad?

3. **Precesión de la órbita de Mercurio**: Observaciones astronómicas realizadas en en el siglo XIX mostraban que el perihelio de Mercurio que es el punto más cercano al sol no era estático, por el contrario, rota lentamente alrededor del sol. La influencia de los otros planetas no respondían completamente al valor de precesión observado y el perihelio de Mercurio avanza con una velocidad de precesión cercana a 43 segundos de arco por siglo. Fue la Relatividad General de Albert Einstein la teoría que dio una respuesta satisfactoria a dicho fenómeno. Esta teoría predice un corrección a la ley de gravitación de Newton dada por:

$$\vec{F} = -\frac{GM_1M_2}{r^2}\left(1+\frac{\alpha}{r^2}\right)\hat{r}. \tag{4.20}$$

   Para Mercurio $\alpha = 1{,}1 \times 10^{-8}\ au^2$. Dado el valor pequeño de α se requiere un alto nivel de precisión para la integración de la órbita en cada revolución. Para encontrar dicho resultado se requiere una ligera modificación al método de Verlet:

$$\begin{aligned} \vec{r}(t+\Delta t) &= \vec{r}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}(t)(\Delta t)^2 \\ \vec{v}(t+\Delta t) &= \vec{v}(t) + \frac{\Delta t}{2}(\vec{a}(t+\Delta t) + \vec{a}(t)) \end{aligned} \tag{4.21}$$

Los parámetros de órbita para Mercurio son: $e = 0{,}205630$ y $a = 0{,}387098$. Inicializar la órbita en el afelio con:

$$\vec{r}(0) = [a(1+e), 0.]$$
$$\vec{v}(0) = [0., \sqrt{G\frac{(1-e)}{a(1+e)}}] \tag{4.22}$$

Demuestre las expresiones de la Ecuaciones (4.21) y (4.22). Usar un paso temporal del mismo orden de $\alpha \sim \Delta t$ para poder medir el efecto de la precesión. Genere al menos 10 órbitas alrededor del Sol guardando en un archivo de datos el tiempo en años terrestres que tarda mercurio en llegar al perihelio y el ángulo de llegada que debe ser cercano a 180°. Usando el archivo de datos gráfique 2 veces el ángulo de llegada vs 2 veces el tiempo en años terrestre que le toma en llegar de nuevo al perihelio. La gráfica que se obtiene es: (*Hint: el perihelio es fácil de encontrar dado que es el radio mínimo a lo largo de la óribta, Be careful con la estrategia para calcular ángulo, además de convertir los grados por año a segundos de arco por siglo*)

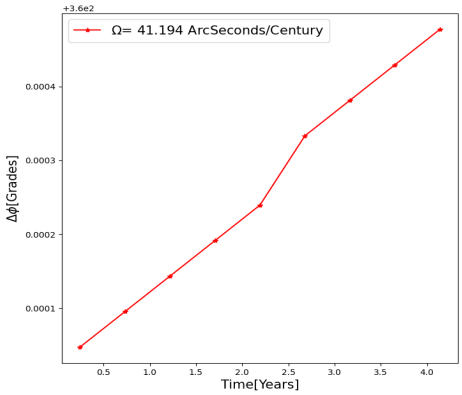

Figura 4.1: Ángulo que se desplaza el perihelio de Mercurio a lo largo de su movimiento alrededor del sol. Se han realizado 10 revoluciones para encontrar la velocidad de precesión, que tiene un error relativo $\approx 2\,\%$ con respecto al valor observado.

Capítulo 5

Series de Fourier

5.1. Periodicidad de una función

Una función periódica se puede definir como una función para la cuál:

$$f(t) = f(t+\tau), \qquad \text{para todo} \qquad t \tag{5.1}$$

la constante que satisface esta condición se denomina periodo de la función, en general:

$$f(t) = f(t+n\tau), \qquad n = 0, \pm 1, \pm 2, ... \tag{5.2}$$

por ejemplo, si $f(t) = cos(\omega_1 T) + cos(\omega_2 T)$ es periódica es posible encontrar m y n tales que:

$$\begin{aligned} \omega_1 T &= 2\pi m \\ \omega_2 T &= 2\pi n \end{aligned} \tag{5.3}$$

Asignando algunos valores se tiene:

$$cos(1/3(t+\tau)) + cos(1/4(t+\tau)) = cos(t/3) + cos(t/4) \tag{5.4}$$

como $cos(\theta + 2\pi m) = cos(\theta)$, se evidencia por ensayo y error que $m = 4$ y $n = 3$, por tanto, $T = 24\pi$.

Sea $f(t)$ una función periódica de periódo T, $f(t)$ puede ser representada por la serie trigonométrica:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n cos(n\omega_0 t) + b_n sin(n\omega_0 t)) \tag{5.5}$$

donde ω_0 es la componente fundamental y $w_n = n\omega_0$ es la enésima armónica. Para entender este resultados regresemos al concepto de polinomios ortonormales. Sea $\phi_k(t)$ un conjunto de funciones ortonormales es un intervalo $a < t < b$ que cumplan:

$$\int_a^b \phi_m(t)\phi_n(t)dt = \begin{cases} 0 & m \neq n \\ r_n & m = n \end{cases} \tag{5.6}$$

Para el conjunto particular de las funciones sinusoidales:

$$\int_{-T/2}^{T/2} cos(m\omega_0 t)cos(n\omega_0 t)dt = \begin{cases} 0 & m \neq n \\ T/2 & m = n \neq 0 \end{cases} \tag{5.7}$$

$$\int_{-T/2}^{T/2} sin(m\omega_0 t)sin(n\omega_0 t)dt = \begin{cases} 0 & m \neq n \\ T/2 & m = n \neq 0 \end{cases} \tag{5.8}$$

$$\int_{-T/2}^{T/2} sin(m\omega_0 t)cos(n\omega_0 t)dt = 0 \qquad \forall m, n \tag{5.9}$$

Podemos proyectar sobre uno de los conjunto base y usar las condiciones de ortogonalidad para encontrar los coeficientes de la expansión.

$$\begin{aligned} \int_{-T/2}^{T/2} f(t)sin(m\omega_0 t)dt &= \frac{1}{2}a_0 \int_{-T/2}^{T/2} sin(m\omega_0 t)dt \\ &+ \sum_{n=1}^{\infty} a_n \int_{-T/2}^{T/2} cos(n\omega_0 t)sin(m\omega_0 t)dt \\ &+ \sum_{n=1}^{\infty} b_n \int_{-T/2}^{T/2} sin(n\omega_0 t)sin(m\omega_0 t)dt \\ &= \sum_{n=1}^{\infty} b_n \frac{T}{2}\delta_{nm} \end{aligned} \tag{5.10}$$

por tanto:

$$b_m = \frac{2}{T}\int_{-T/2}^{T/2} f(t)sin(m\omega_0 t)dt \tag{5.11}$$

proyectando sobre el otro conjunto base:

$$\begin{aligned} \int_{-T/2}^{T/2} f(t)cos(m\omega_0 t)dt &= \frac{1}{2}a_0 \int_{-T/2}^{T/2} cos(m\omega_0 t)dt \\ &+ \sum_{n=1}^{\infty} a_n \int_{-T/2}^{T/2} cos(n\omega_0 t)cos(m\omega_0 t)dt \\ &+ \sum_{n=1}^{\infty} b_n \int_{-T/2}^{T/2} sin(n\omega_0 t)cos(m\omega_0 t)dt \\ &= \sum_{n=1}^{\infty} a_n \frac{T}{2}\delta_{nm} \end{aligned} \tag{5.12}$$

por tanto:

$$a_m = \frac{2}{T}\int_{-T/2}^{T/2} f(t)cos(m\omega_0 t)dt \tag{5.13}$$

Finalmente, a_0 resulta ser el valor medio de la función en un periodo.

$$a_0 = \frac{2}{T}\int_{-T/2}^{T/2} f(t)dt \tag{5.14}$$

Ejemplo: Calcular la suma $\sum_{n=1}^{\infty}\frac{1}{n^6}$ usando la serie de Fourier de la función t^2 en el intervalo $-\pi < t < \pi$.

La serie de Fourier de $f(t) = t^2$ con la condición $f(t+2\pi) = f(t)$ es:

$$t^2 = \frac{\pi^2}{3} + 4\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}cos(nt). \tag{5.15}$$

integrando la serie se tiene:

$$\frac{1}{12}t(t^2-\pi^2) = \sum_{n=1}^{\infty}\frac{(-1)^n}{n^3}sin(nt) \tag{5.16}$$

Tengamos en cuenta la identidad de Parseval:

$$\frac{1}{T}\int_{-T/2}^{T/2}[f(t)]^2dt = \frac{1}{4}a_0^2 + \frac{1}{2}\sum_{n=1}^{\infty}(a_n^2+b_n^2). \tag{5.17}$$

Notemos que la función t^3 es impar de modo que los coeficientes diferentes de cero son:

$$b_n = \frac{(-1)^n}{n^3} \tag{5.18}$$

usando la identidad de Parseval:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}(\frac{1}{12}t(t^2-\pi^2))^2dt = \frac{1}{2}\sum_{n=1}^{\infty}\frac{(-1)^{2n}}{n^6} \tag{5.19}$$

realizando los pasos intermedios se obtiene el valor de la suma

$$\sum_{n=1}^{\infty}\frac{1}{n^6} = \frac{\pi^6}{945} \tag{5.20}$$

5.2. Transformada discreta de Fourier

En el caso de que la función sea no periódica es posible expandir estas ideas a una transformada de Fourier. Dado que no existe la periodicidad las frecuencias deben entenderse en términos de la variabilidad de la función $f(t)$. Los coeficientes de la transformada discreta de Fourier están dados por:

$$A_k = \sum_{m=0}^{N-1} A_m e^{-2\pi jmk/N}, \qquad k = 0, 1, ..., N-1. \tag{5.21}$$

donde A_m son los puntos muestrales de $f(t)$, de modo que se necesita extraer una muestra finita de la función cuya transformada queremos calcular. La función inversa está dada por:

$$A_m = \frac{1}{N}\sum_{k=0}^{N-1} A_k e^{2\pi jmk/N}, \qquad m = 0, 1, ..., N-1. \tag{5.22}$$

donde A_k son los coeficientes de Fourier de $f(t)$. Expandiendo la primer suma podemos entender como implementar el algoritmo.

$$A_k = A_0 e^{-2\pi j0k/N} + A_1 e^{-2\pi j1k/N}, ..., A_{N-1} e^{-2\pi j(N-1)k/N} \tag{5.23}$$

5.2.1. Condiciones de Dirichlet

a) $f(t)$ debe ser absolutamente integrable $\int_{-\infty}^{\infty} f(t)dt = M$.

b) En cualquier intervalo finito debe existir un número finito de máximos y mínimos.

c) En cualquier intervalo finito debe existir un número finito de discontinuidades.

5.2.2. Transformada de Fourier 2D

Es posible generalizar la transformada discreta de Fourier al caso 2D. Particular $f(x, y)$ puede ser una imagen que puede ser procesada por dicho algoritmo.

$$F(U, V) = \frac{1}{NM}\sum_{x=0}^{N-1}\sum_{y=0}^{M-1} f(x, y)e^{-2\pi j(xU/N+yV/M)} \tag{5.24}$$

cuya transformada inversa está dado por:

$$f(x, y) = \sum_{U=0}^{N-1}\sum_{V=0}^{M-1} F(U, V)e^{2\pi j(xU/N+yV/M)} \tag{5.25}$$

5.2.3. Filtros

En el espacio reciproco es posible por medio de la operación de convolución aplicar filtros de baja y alta frecuencia. Un filtro de baja frecuencia está dado por:

$$H(U, V) = \begin{cases} 1 & F(U, V) \leq R_0 \\ 0 & \text{En otro caso} \end{cases} \tag{5.26}$$

Por otro lado, un filtro de alta frecuencia está dado por:

$$H(U, V) = \begin{cases} 1 & F(U, V) \geq R_0 \\ 0 & \text{En otro caso} \end{cases} \tag{5.27}$$

Matemáticamente, filtrar una imagen, un archivo de audio o un conjunto de datos, se refiere a una convolución entre la transformada de Fourier y un determinado filtro de la siguiente manera:

a) Hallar $F(U, V)$.

b) Filtrar usando $G(U, V) = H(U, V) * F(U, N)$.

c) Hallar la transformada inversa $G^{-1}(U, V)$.

d) La imagen final es $G(x, y)$.

5.3. Ejercicios

1. Demostrar las ecuaciones (5.7),(5.8) y (5.9).

Capítulo 6

Ecuaciones Diferenciales Ordinarias

6.1. Método de Euler

Este método consiste en desarrollar un algoritmo numérico para resover el problema de valores iniciales:

$$y'(x) = f(x, y), \qquad y(a) = y_a, \qquad x \in [a, b], \tag{6.1}$$

siendo $f(x, y)$ una función acotada y continua en la variable x. Consideremo el dominio $[a, b]$ discretizando en $n + 1$ puntos equiespaciados.

$$x_i = x_0 + ih, \qquad i = 0, 1, ..., n \qquad x_0 = a. \tag{6.2}$$

El paso de la discretización es por tanto:

$$h = \frac{b - a}{n} \tag{6.3}$$

Para encontrar una expresión discreta, comencemos con el desarrollo en series de la función $y(x)$, en algun punto de la partición x_{i+1}:

$$y(x_{i+1}) = y(x_i) + y'(x_i)(x_{i+1} - x_i) + \frac{y''(x_i)}{2}(x_{i+1} - x_i)^2 + ... \tag{6.4}$$

Usando la Ecuación (6.2). Se obtiene:

$$y(x_{i+1}) = y(x_i) + hy'(x_i) + \mathcal{O}(h^2) \tag{6.5}$$

Si despejamos el operador derivada y restamos ambos miembros por $f(x, y)$:

$$y'(x) - f(x, y) = \frac{y(x_{i+1}) - y(x_i)}{h} - f(x_i, y_i) - \mathcal{O}(h^2) \tag{6.6}$$

El último término se refiere el truncamiento del algoritmo, de modo que el método de Euler es de orden 1 en aproximación. Finalmente se obtiene la siguiente forma iterativa:

$$y_{i+1} = y_i + hf(x_i, y_i), \qquad i = 0, 1, ...n \qquad y(a) = y_a. \tag{6.7}$$

Cabe resaltar que el método es consistente dado que al tender el tamaño de la discretización ($h \to 0$) a cero, error local tiende a cero.

$$\mathcal{O}(h) \to 0 \text{ cuando } h \to 0. \tag{6.8}$$

6.2. Método de Euler mejorado

Este método consiste en mejorar la aproximación a orden 1 usando la regla de integración del trapecio. Tomemos el problema de valores iniciales:

$$y'(x) = f(x, y), \qquad y(x_0) = y_0 \tag{6.9}$$

Integramos en un paso de integración desde x_0 hasta $x_1 = x_0 + h$.

$$y(x_1) - y(x_0) = \int_{x_0}^{x_1} f(x, y)dx \tag{6.10}$$

Para calcular el miembro derecho usamos la regla del trapecio simple.

$$\int_{x_0}^{x_1} f(x, y)dx = \frac{h}{2}(f(x_0, y_0) + f(x_1, y_1)) \tag{6.11}$$

La cuestión es encontrar el valor de y_1. Para tal proposito usamos el método de Euler $y_1 = y_0 + hf(x_0, y_0)$. Entonces tenemos:

$$y_1 = y_0 + \frac{h}{2}(f(x_0, y_0) + f(x_0 + h, y_0 + hf(x_0, y_0)) \tag{6.12}$$

Para la implementación computación se definen las siguiente constantes y regla de iteración:

$$\begin{aligned} k_1 &= f(x_0, y_0) \\ k_2 &= f(x_0 + h, y_0 + hf(x_0, y_0)) \\ y_1 &= y_0 + \frac{h}{2}(k_1 + k_2) \end{aligned} \tag{6.13}$$

6.3. Métodos de Runge-Kutta

La mejora del método siempre es posible teniendo en cuenta más términos en la estimación de las pendientes. En general, esta familia de métodos de un paso se denominan métodos de Runge-Kutta para la aproximación de ecuaciones diferenciales ordinarias. La aproximación de orden 3 está dada por:

$$\begin{aligned} k_1 &= f(y_n, x_n) \\ k_2 &= f(y_n + \frac{1}{2}k_1, x_n + \frac{h}{2}) \\ k_3 &= f(y_n - k_1 h + 2k_2 h, x_n + h) \\ y_{n+1} &= y_n + \frac{h}{6}(k_1 + 4k_2 + k_3) \end{aligned} \tag{6.14}$$

La aproximación de orden 4 está dada por:

$$\begin{aligned} k_1 &= f(y_n, x_n) \\ k_2 &= f(y_n + \frac{1}{2}k_1 h, x_n + \frac{h}{2}) \\ k_3 &= f(y_n + \frac{1}{2}k_2 h, x_n + \frac{h}{2}) \\ k_4 &= f(y_n + k_3 h, x_n + h) \\ y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \end{aligned} \tag{6.15}$$

6.4. Métodos Multi-paso

La consistencia de los métodos de un paso (familia de métodos Runge-Kutta) representan el primer acercamiento numérico a la solución de problemas de valor inicial. Sin embargo, aunque los métodos son consistentes y convergentes cuando aumentamos el orden de la aproximación, tienen la limitación que solo se usan la información del paso presente para estimar el valor de las funciones en el siguiente paso (futuro). Los métodos Multi-paso usan la información del paso presente y la información acumulada en los pasos anteriores para mejorar la estimación del siguiente valor. Estos métodos se denominan métodos de Adams, los cuales se clasifican en métodos explícitos (Adams-Bashforth) e implícitos (Adams-Moulton). Los métodos explícitos (predictores) usan la información acumulada hasta el paso presente para estimar las funciones en el siguiente paso. Por otro lado, los métodos implícitos (correctores) usan la información acumulada en el paso presente más la predicción realizada explícitamente, para realizar una corrección del siguiente paso. En general, es necesario usar ambos métodos en conjunto para mejorar el cálculo de las funciones en un esquema de integración denominado predictor-corrector approach.

El método multi-paso lineal más general posible es:

$$y_{n+1} = \psi_k(t_0, y_0, ..., t_{n+1}, y_{n+1}), \tag{6.16}$$

donde ψ_k es una transformación lineal en las y_k y evaluada en esos puntos. No obstante, el problema de valor inicial solo nos brinda el primer valor de la función y_0. Un esquema requiere k puntos para iniciar el esquema multi-paso. Convencionalmente se utiliza métodos Runge-Kutta de orden 4 para calcular esos k puntos iniciales. De esta manera, conocemos los primeros k valores de la función $\{y_0, y_1, ..., y_l\}$ en $\{t_0, t_1, ..., t_l\}$ con $l = 0, 1, ..., k$. Con esta información calculamos el polinomio interpolador de orden k:

$$p_k(x) = \sum_{i=0}^{k} f(x_i)\mathcal{L}_i(x), \tag{6.17}$$

$$\mathcal{L}_i(x) = \prod_{j=0, j\neq i}^{k} \frac{x - x_j}{x_i - x_j} \tag{6.18}$$

usando las bases cardinales se escribe el problema de valor inicial como:

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} p_k(x)dx = y_n + \sum_{i=0}^{k} f_i \int_{t_n}^{t_{n+1}} \mathcal{L}_i(x)dx \tag{6.19}$$

Definimos:

$$\beta_i = \int_{t_n}^{t_{n+1}} \mathcal{L}_i(x)dx \tag{6.20}$$

De este modo, el problema de valor inicial se reduce a un problema de cuadraturas de orden k:

$$y_{n+1} = y_n + \sum_{i=0}^{k} \beta_i f_i \tag{6.21}$$

6.4.1. Adams-Bashforth de 2 puntos

Supongamos que tenemos $\{f_n, t_n\}$ y $\{f_{n-1}, t_{n-1}\}$ con $t_n - t_{n-k} = kh$. El polinomio interpolador que pasa por dichos puntos es:

$$p_1(t) = \frac{t - t_{n-1}}{t_n - t_{n-1}} f_n + \frac{t - t_n}{t_{n-1} - t_n} f_{n-1} \tag{6.22}$$

De modo que el método iterativo de 2 puntos es:

$$y_{n+1} = y_n + \frac{f_n}{h} \int_{t_n}^{t_{n+1}} (t - t_{n-1})dt - \frac{f_{n-1}}{h} \int_{t_n}^{t_{n+1}} (t - t_n)dt \tag{6.23}$$

$$y_{n+1} = y_n + \frac{h}{2}(3f_n - f_{n-1}) \tag{6.24}$$

Finalmente, recordando el error en la estimación de un polinomio interpolante, el error local para dos puntos es $\mathcal{O}(h^2)$.

6.4.2. Adams-Moulton de 2 puntos

Usando la predicción (y_{n+1}) estimada por el método anterior, tenemos $\{f_{n+1}, t_{n+1}\}$ y $\{f_n, t_n\}$ con $t_n - t_{n-k} = kh$. Por tanto, el polinomio interpolador es:

$$p_1(t) = \frac{t - t_n}{t_{n+1} - t_n} f_{n+1} + \frac{t - t_{n+1}}{t_n - t_{n+1}} f_n \tag{6.25}$$

De modo que el método implícito está dado por:

$$y_{n+1} = y_n + \frac{f_{n+1}}{h}\int_{t_n}^{t_{n+1}}(t-t_n)dt - \frac{f_n}{h}\int_{t_n}^{t_{n+1}}(t-t_{n+1})dt \tag{6.26}$$

$$y_{n+1} = y_n + \frac{h}{2}(f_{n+1} - f_n) \tag{6.27}$$

Esto corrige el valor de la función calculada por el método explícito con un error local $\mathcal{O}(h^2)$. *Note que este resultada es simplemente el método trapezoidal.*

En general, para mejorar la precisión significativamente se usa un método corrector que tenga una precisión al orden siguiente ($\mathcal{O}(h^4)$), por tanto, para una formula de dos puntos se tiene un método de Adams-Moulton dado por:

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}) \tag{6.28}$$

6.5. Evolución de un pandemia, Covid-19

Se puede modelar la evolución de una pandemia en término de ecuaciones diferenciales teniendo en cuenta las siguientes variables: S es el número de personas susceptibles a contraer la enfermedad, I es el número de personas que han adquirido la enfermedad y son capaces de transmitirla. Adicionalmente, el modelo requiere dos parámetros que caractericen a la población (de N individuos). β es la tasa de infección y γ es la tasa de recuperación, las cuales consideramos constantes, al igual que la población. Es posible representar la evolución de la enfermedad a través del siguiente sistema de ecuaciones diferenciales.

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \end{aligned} \tag{6.29}$$

para un conjunto de parámetros: $N = 1000$, $S_0 = 1000$, $I_0 = 1$, $\beta = 0{,}002$ y $\gamma = 0{,}5$ se obtiene la siguiente curva de infectados y personas suceptibles:

Podemos hacer una mejora a nuestro modelo dividiendo la población total en k sub-regiones y analizando el flujo de personas entre regiones [1]. Cada región tendrá su propia tasa de infección β_k y la tasa de recuperación γ, la cuál no depende de la región donde viva el individuo. Adicionalmente, se define el número de reproducción como la tasa de infección entre la capacidad de recuperación por región:

$$R_k = \frac{\beta_k}{\gamma} \tag{6.30}$$

Usando esta información se escribe el siguiente sistema de ecuaciones:

$$\begin{aligned} \frac{dS_k}{dt} &= -\frac{\beta_k I_k}{N_k} S_k \\ \frac{dI_k}{dt} &= \frac{\beta_k S_k}{N_k} I_k - \gamma I_k \\ \frac{dR_k}{dt} &= \gamma I_k \end{aligned} \tag{6.31}$$

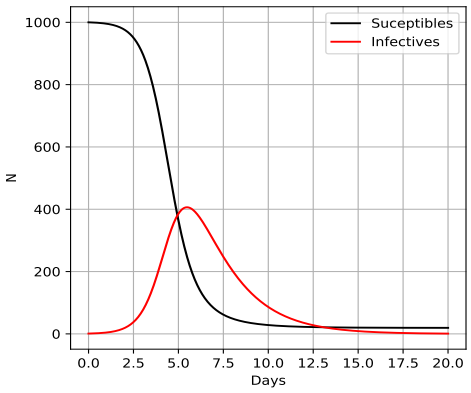

Figura 6.1: Modelo de propagación de una enfermedad a lo largo del tiempo. Notar como aparece un máximo en el número de personas infectadas conocida como: *pico de la pandemia*.

Durante la evolución se tiene la siguiente restricción $N_k = S_k + I_k + R_k$. Para determinar unívocamente la solución de las ecuaciones en cada sub-región debemos entender como se transportan las personas de una región a otra. Se puede entender dicho transporten en términos de una matriz de transferencia denominada genéricamente *origin-destination matrix*.

$$S_k(t) = T_{kj} S_j(t), \tag{6.32}$$

donde $T_{kj} = \mathbb{I} + \alpha(F^{in}_{kj} + F^{out}_{kj})$. Las matrices $\mathbb{F}$ controlan el flujo de entrada y salida entre las regiones j y k y α es el parámetro de movilidad. Veamos el algoritmo a través de un ejemplo: consideremos tres regiones con el siguiente número de individuos.

$$N = \begin{pmatrix} 50 \\ 20 \\ 30 \end{pmatrix} \tag{6.33}$$

En un momento determinado supongamos que de la región 1, salen 10 personas a la región 2 y 20 personas a la región 3. de la región 2, salen 5 personas a la región 1 y 7 personas a la región 3. Finalmente, de la región 3, salen 8 personas a la región 1 y 12 personas a la región 2. Notar que la matriz de salida es diagonal, puesto que:

$$\begin{pmatrix} -30 \\ -12 \\ -20 \end{pmatrix} = \begin{pmatrix} -30 & 0 & 0 \\ 0 & -12 & 0 \\ 0 & 0 & -20 \end{pmatrix} = \begin{pmatrix} -3/5 & 0 & 0 \\ 0 & -3/5 & 0 \\ 0 & 0 & -2/3 \end{pmatrix} \begin{pmatrix} 50 \\ 20 \\ 30 \end{pmatrix} = F^{out}_{kk} S_k \tag{6.34}$$

La matriz de entrada resulta menos intuitiva. Note que en la región 1 entran 5 personas que provienen de la región 2 y 8 personas de la región 3. Entonces:

$$\begin{aligned} N_1 &= F_{12}^{in} N_2 \\ +5 &= F_{12} 20 \\ N_1 &= F_{13}^{in} N_3 \\ +8 &= F_{13} 30 \end{aligned} \tag{6.35}$$

Realizando el mismo análisis para las otras regiones obtenemos las matriz de entrada.

$$F^{in} = \begin{pmatrix} 0 & 1/4 & 4/15 \\ 1/5 & 0 & 2/5 \\ 2/5 & 7/20 & 0 \end{pmatrix} \tag{6.36}$$

Efectivamente la suma de estos operadores debe generar el flujo neto de personas en cada región:

$$\Phi = \begin{pmatrix} -3/5 & 1/4 & 4/15 \\ 1/5 & -3/5 & 2/5 \\ 2/5 & 7/20 & -2/3 \end{pmatrix} \begin{pmatrix} 50 \\ 20 \\ 30 \end{pmatrix} = \begin{pmatrix} -17 \\ 10 \\ 7 \end{pmatrix} \tag{6.37}$$

De modo que la matriz de origen-destino con α es efectivamente:

$$\begin{pmatrix} 33 \\ 30 \\ 37 \end{pmatrix} = (\mathbb{I} + \Phi) \begin{pmatrix} 50 \\ 20 \\ 30 \end{pmatrix} \tag{6.38}$$

$$\mathbb{T} = \begin{pmatrix} 2/5 & 1/4 & 4/15 \\ 1/5 & 2/5 & 2/5 \\ 2/5 & 7/20 & 1/3 \end{pmatrix} \tag{6.39}$$

cuando las personas regresan a casa la matriz de transporte sera $\mathbb{T}^{-1} = (\mathbb{I} + \Phi)^{-1}$. Variando el valor de α, es decir controlando la movilidad de las personas es posible integrar las ecuaciones diferenciales y obtener un comportamiento más realista de la propagación de la enfermedad. Adicionalmente, los valores de la matriz de trasnferencia deben entenderse como la probabilidad (sumando los valores de cada columna) de que una persona decida entrar o salir de una región específica.

6.6. Sistemas de ecuaciones diferenciales con Runge-Kutta

Para resolver sistemas de ecuaciones diferenciales por medio del método de Runge-Kutta, definimos vectorialmente el problema de valor inicial de N ecuaciones diferenciales:

$$\dot{y}_n(t) = f_n(y_1, y_2, ..., y_N, t), \qquad n = 1, 2, ..., N. \tag{6.40}$$

El algoritmo se puede resumir como sigue:

1. Dar $y_n = y_n(0)$ para $t = 0$ y $n = 1, 2, ..., N$.

2. Evaluar $k_n^{(1)} = h f_n(y_1, y_2, ..., y_N, t)$ para $n = 1, 2, ..., N$.

3. Evaluar $k_n^{(2)} = hf_n(y_1 + \frac{k_1^{(1)}}{2}, y_2 + \frac{k_2^{(1)}}{2}, ..., y_N + \frac{k_N^{(1)}}{2}, t + \frac{h}{2})$ para $n = 1, 2, ..., N$.
4. Evaluar $k_n^{(3)} = hf_n(y_1 + \frac{k_1^{(2)}}{2}, y_2 + \frac{k_2^{(2)}}{2}, ..., y_N + \frac{k_N^{(2)}}{2}, t + \frac{h}{2})$ para $n = 1, 2, ..., N$.
5. Evaluar $k_n^{(4)} = hf_n(y_1 + k_1^{(3)}, y_2 + k_2^{(3)}, ..., y_N + k_N^{(3)}, t + h)$ para $n = 1, 2, ..., N$.
6. $y_n(t + h) = y_n(t) + \frac{1}{6}[k_n^{(1)} + 2k_n^{(2)} + 2k_n^{(3)} + k_n^{(4)}]$.
7. $t = t + h$. Ir al paso (2).

6.7. Ejercicios

1. Demuestre las expresiones de los métodos de Runge-Kutta de orden 3 (6.14) y orden 4 (6.15).
2. Demuestre la formula de iteración para el método de Adams-Bashforth de tres puntos.

$$y_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}) \tag{6.41}$$

Hint: Integrar el polinomio interpolador para los puntos: $\Omega = \{(t_{n-2}, f_{n-2}), (t_{n-1}, f_{n-1}), (t_n, f_n)\}$
Recuerde usar esta formula con un método corrector al siguiente orden:

$$y_{n+1} = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) \tag{6.42}$$

3. Demuestre la formula de iteración para el método de Adams-Moulton de tres puntos.

$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}) \tag{6.43}$$

Hint: Integrar el polinomio interpolador para los puntos: $\Omega = \{(t_{n-1}, f_{n-1}), (t_n, f_n), (t_{n+1}, f_{n+1})\}$

Capítulo 7

Ecuaciones Diferenciales Parciales

Una ecuación diferencial en derivadas parciales (EDP) es una relación matemática entre las derivadas parciales de una función desconocida y las variables independientes. El orden la derivada más alta define el orden de la EDP. Una EDP se considera lineal si es de primer grado en la variable dependiente y en su derivada parcial. Si cada término de la PDE contiene o a la variable dependiente o una de sus derivadas, la EDP es llamada homogénea, de lo contrario es no homogénea. La forma más general de una ecuación diferencial es:

$$a\partial^2_{xx}u + b\partial^2_{xy}u + c\partial^2_{yy} + d\partial_x u + e\partial_y u + fu = g \tag{7.1}$$

El discriminante de la ecuación diferencial las clasifica en tres familias: elíptica, parabólica e hiperbólicas. Estos nombres están en analogía con las secciones cónicas.

$$b^2 - 4ac = \begin{cases} < 0 & \text{Elíptica} \\ = 0 & \text{Parabólica} \\ > 0 & \text{Hiperbólica} \end{cases} \tag{7.2}$$

Algunas de las condiciones de frontera más usadas son:

1. Condiciones de Dirichlet:
$$y(x) = f(x), \qquad \forall_x \in \partial\Omega \tag{7.3}$$

2. Condiciones de Von Neummann.
$$\frac{\partial y(x)}{\partial \hat{n}} = \nabla y(x) \cdot \hat{n}(x) \tag{7.4}$$

   donde $\hat{n}$ es el vector normal en la frontera y $f(x)$ una función escalar.

7.1. Problemas elípticos

Estas EDP describen procesos de equilibrio independientes del tiempo, en ese sentido son problemas de condiciones de frontera por ejemplo: campos de temperatura, fluidos incompresibles o calculo del potenciales.

Ecuación de Laplace 2D:

Por excelencia uno de los problemas más comunes y útiles es la ecuación de Laplace, la cuál tiene la siguiente forma:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{7.5}$$

Discretizando el dominio solución $(x, y) \in \mathbb{R} = [a, b] \times [c, d]$, tenemos la siguiente notación: $u(x_i, y_j) = u_{i,j}$, donde $i = 1, 2, ..., nx - 1$ y $j = 1, 2, ..., ny - 1$. En diferencias finitas, los operadores diferenciales tienen la siguiente forma:

$$\frac{u_{i+1,i} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} + \frac{u_{i,i+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta y)^2} = 0 \tag{7.6}$$

comúnmente se ajustan los pasos en ambas direcciones como iguales, lo que permite escribir la siguiente ecuación iterativa para el potencial:

$$u_{i,j} = \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1}, u_{i,j-1}}{4} \tag{7.7}$$

Este tipo de ecuaciones solo tiene condiciones de frontera dadas por:

$$\begin{aligned} u(a, y) &= h_1(y) \\ u(b, y) &= h_2(y) \\ u(x, c) &= h_3(x) \\ u(x, d) &= h_4(x) \end{aligned} \tag{7.8}$$

Veamos como este tipo de problema conduce a un sistema de ecuaciones lineales: Tomemos un región solución dada por $A = 40 \times 40\ cm^2$ con $\Delta x = \Delta y = 10\ cm$. Las temperaturas en grados Celsius en las fronteras son:

$$\begin{aligned} u(0, y) &= 75 \\ u(40, y) &= 50 \\ u(x, c) &= 100 \\ u(x, d) &= 0 \end{aligned} \tag{7.9}$$

Utilizando la ecuación iterativa llegamos al siguiente problema matricial.

$$\begin{pmatrix} -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -4 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & -4 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 \end{pmatrix} \begin{pmatrix} u_{22} \\ u_{23} \\ u_{24} \\ u_{32} \\ u_{33} \\ u_{34} \\ u_{42} \\ u_{43} \\ u_{44} \end{pmatrix} = \begin{pmatrix} -175 \\ -75 \\ -75 \\ -100 \\ 0 \\ 0 \\ -150 \\ -50 \\ -50 \end{pmatrix} \tag{7.10}$$

La resolución de este problema nos brinda el valor de temperatura o potencial en cada uno de esos 9 puntos. Sin embargo, si cambiamos la discretización y los valores frontera resulta poco practico la solución de los sistemas lineales vistos hasta el momento. Por tal razón, se introducen los métodos de relajación para solucionar de manera optima este tipo de ecuaciones diferenciales parciales.

7.1.1. Método de relajación sucesiva

Estos métodos realizan un ponderación de la solución de un sistema lineal en una iteración específica. La expresión general para un peso w dado es:

$$x_i^j = wx_i^j + (1-w)x_i^{j-1} \tag{7.11}$$

donde el índice i se relaciona al valor incógnita y j al valor de la iteración. Dependiendo del valor del peso, se tiene la siguiente calificación.

- Sub-relajación $0 < w < 1$. No convergente (para movimiento de fluidos puede converger).
- No modificado $w = 1$.
- Sobre-relación $1 < w < 2$. Acelera convergencia.
- Divergente $w > 2$.

Encontrar el valor adecuado del peso dada una discretización es un asunto complicado del álgebra matricial numérica. Por ejemplo, para una discretización cuadrada de N puntos, el peso optimo está dado por:

$$w = \frac{2}{1+\frac{\pi}{N}} \tag{7.12}$$

Ejemplo:

Encontrar x_3^1 usando el método de sobre-relajación del siguiente sistema de ecuaciones lineales. Considere $w = 1{,}25$. y $\vec{x}_0 = (0,0,0)$.

$$\begin{aligned} 4x_1 - x_2 &= 0 \\ -x_1 + 4x_2 - x_3 &= 6 \\ -x_2 + 4x_3 &= 2 \end{aligned} \tag{7.13}$$

Usando el acercamiento por Gauss-Seidel tenemos:

$$\begin{aligned} x_1^1 &= \frac{2+x_2^0}{4} = \frac{2+0}{4} = 0{,}5 \\ x_2^1 &= \frac{6+x_1^1+x_3^0}{4} = \frac{6+0{,}5+0}{4} = 1{,}625 \\ x_3^1 &= \frac{2+x_2^1}{4} = \frac{2+1{,}625}{4} = 0{,}906 \end{aligned} \tag{7.14}$$

Ahora usemos sobre-relajación (para mostrar como funciona).

$$\begin{aligned}
x_1^1 &= \frac{2+x_2^0}{4} = \frac{2+0}{4} = 0{,}5 \\
x_1^1 &= (1{,}25)(0{,}5) + (1-1{,}25)(0) = 0{,}625 \\
x_2^1 &= \frac{6+x_1^1+x_3^0}{4} = \frac{6+0{,}625+0}{4} = 1{,}656 \\
x_2^1 &= (1{,}25)(1{,}656) + (1-1{,}25)(0) = 2{,}070 \\
x_3^1 &= \frac{2+x_2^1}{4} = \frac{2+2{,}070}{4} = 1{,}017 \\
x_3^1 &= (1{,}25)(1{,}017) + (1-1{,}25)(0) = 1{,}271
\end{aligned} \tag{7.15}$$

Aunque la solución exacta para $x_3 = 27/28$, el método de sobre-relajación parece más distantes de la solución exacta. Sin embargo, se deja al lector la implementación de ambas rutinas para verificar la potencial del método de relajación sucesiva. Adicionalmente, es posible imponer un criterio de convergencia del método:

$$\left| \frac{x_i^j - x_i^{j-1}}{x_i^j} \right| < \epsilon \tag{7.16}$$

donde ϵ es arbitrariamente pequeño. Podemos pensar en la siguiente factorización para la implementación de la relajación sucesiva en ecuaciones diferenciales parciales:

$$u_{i,j}^l = u_{i,j}^{l-1} + w[u_{i,j}^l - u_{i,j}^{l-1}] \tag{7.17}$$

definiendo $r_{i,j} := u_{i,j}^l - u_{i,j}^{l-1}$ se tiene:

$$u_{i,j}^l = u_{i,j}^{l-1} + w r_{i,j} \tag{7.18}$$

Generalmente $r_{i,j} \leq 10^{-10}$.

7.1.2. Operador de Laplace en coordenadas cilíndricas

Es posible extrapolar el método de diferencias finitas a operadores en coordenadas curvilineas, en ese sentido el índice i queda asociado a la coordenada radial ($\rho = i\Delta\rho$) y el indice j está relacionado con la coordenada azimutal ($\phi = j\Delta\phi$). La discretización necesaria para describir el operador de Laplace en coordenadas cilíndricas es:

$$\begin{aligned}
\frac{\partial u}{\partial \rho} &= \frac{u_{i,j} - u_{i-1,j}}{\Delta\rho} \\
\frac{\partial^2 u}{\partial \rho^2} &= \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta\rho)^2} \\
\frac{\partial^2 u}{\partial \phi^2} &= \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{(\Delta\phi)^2}
\end{aligned} \tag{7.19}$$

El operador de Laplace queda definido por:

$$\begin{aligned}\nabla^2 u(\rho,\phi) &= \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{(\Delta\rho)^2} \\ &+ \frac{1}{\rho[i]}\left(\frac{u_{i,j}-u_{i-1,j}}{\Delta\rho}\right) \\ &+ \frac{1}{\rho[i]^2}\left(\frac{u_{i,j+1}-2u_{i,j}+u_{i,j-1}}{(\Delta\phi)^2}\right) \end{aligned} \tag{7.20}$$

multiplicando por $(\Delta\rho)^2$ y definiendo $\lambda = \frac{\Delta\rho}{\Delta\phi}$ tenemos:

$$u_{i+1,j} + u_{i-1,j}\left[1-\frac{\Delta\rho}{\rho[i]}\right] + \frac{\lambda^2}{\rho[i]^2}\left[u_{i,j+1}+u_{i,j-1}\right] = \left[2-\frac{\Delta\rho}{\rho[i]}+\frac{2\lambda^2}{\rho[i]^2}\right]u_{i,j} \tag{7.21}$$

Vamos a definir un coeficiente de escala de estas coordenadas como:

$$c[i] := 2-\frac{\Delta\rho}{\rho[i]}+\frac{2\lambda^2}{\rho[i]^2} \tag{7.22}$$

Finalmente tenemos una expresión implícita para este operador:

$$u_{i,j} = \frac{1}{c[i]}\left[u_{i+1,j}+u_{i-1,j}\left[1-\frac{\Delta\rho}{\rho[i]}\right]+\frac{\lambda^2}{\rho[i]^2}\left[u_{i,j+1}+u_{i,j-1}\right]\right] \tag{7.23}$$

como todos los problemas con simetría cilíndrica se requiere la continuidad del potencial para el ángulo azimutal: $u(\rho,\phi+2\pi) = u(\rho,\phi)$.

7.1.3. Operador de Laplace en coordenadas esféricas

Para construir este operador diferencial necesitamos las siguientes derivadas:

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2}+\frac{2}{r}\frac{\partial u}{\partial r}+\frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2}+\frac{1}{r^2 tan(\theta)}\frac{\partial u}{\partial\theta}+\frac{1}{r^2 sin^2(\theta)}\frac{\partial^2 u}{\partial\phi^2} \tag{7.24}$$

El índice i está asociado a la coordenada radial ($r = i\Delta r$) y el indice j está relacionado a la coordenada polar ($\theta = i\Delta\theta$) y el índice k está relacionado con la coordenada azimutal ($\phi = j\Delta\phi$). El operador Laplaciano en diferencias finitas queda dado por:

$$\begin{aligned}\nabla^2 u &= \frac{u_{i+1,j,k}-2u_{i,j,k}+u_{i-1,j,k}}{(\Delta r)^2} \\ &+ \frac{2}{r}\left[\frac{u_{i+1,j,k}+u_{i-1,j,k}}{2\Delta r}\right] \\ &+ \frac{1}{r^2}\left[\frac{u_{i,j+1,k}-2u_{i,j,k}+u_{i,j-1,k}}{(\Delta\theta)^2}\right] \\ &+ \frac{1}{r^2 tan(\theta)}\left[\frac{u_{i,j+1,k}+u_{i,j-1,k}}{2\Delta\theta}\right] \\ &+ \frac{1}{r^2 sin^2(\theta)}\left[\frac{u_{i,j,k+1}-2u_{i,j,k}+u_{i,j,k-1}}{(\Delta\phi)^2}\right] \end{aligned} \tag{7.25}$$

Factorizando $1/(\Delta r)^2$ tenemos la siguiente expresión:

$$\begin{aligned}\nabla^2 u &= \frac{1}{(\Delta r)^2}\Big[u_{i+1,j,k} - 2u_{i,j,k} + u_{i-1,j,k} \\ &+ \frac{\Delta r}{r}\Big[u_{i+1,j,k} + u_{i-1,j,k}\Big] \\ &+ \left(\frac{\Delta r}{r\Delta\theta}\right)^2\Big[u_{i,j+1,k} - 2u_{i,j,k} + u_{i,j-1,k}\Big] \\ &+ \frac{(\Delta r)^2}{(\Delta\theta)r^2 tan(\theta)}\Big[u_{i,j+1,k} + u_{i,j-1,k}\Big] \\ &+ \frac{(\Delta r)^2}{(\Delta\phi)^2 r^2 sin^2(\theta)}\Big[u_{i,j,k+1} - 2u_{i,j,k} + u_{i,j,k-1}\Big]\Big] \end{aligned} \tag{7.26}$$

Definiendo $\lambda := \frac{\Delta r}{\Delta\theta}$ y $\mu := \frac{\Delta r}{\Delta\phi}$. Se puede ver la estrategia de implementación de cada sumando ponderando por los siguientes coeficientes:

$$\begin{aligned} a &= 1 \\ b &= \frac{\Delta r}{r} \\ c &= \left(\frac{\lambda}{r}\right)^2 \\ d &= \left(\frac{\lambda}{r}\right)^2 \frac{\Delta\theta}{tan(\theta)} = c\frac{\Delta\theta}{tan(\theta)} \\ e &= \left(\frac{\mu}{rsin(\theta)}\right)^2 \end{aligned} \tag{7.27}$$

En el caso de implementación de ecuación de onda:

$$\partial_{tt} u(r,\theta,\phi) = \alpha^2 \nabla^2 u(r,\theta,\phi) \tag{7.28}$$

se debe definir un nuevo parámetro $\nu = \frac{\alpha\Delta t}{\Delta r}$, que permita escribir un método explicito para el tiempo.

7.2. Problemas parabólicos

Las EDP parabólicas describen procesos físicos que están cambiando lentamente en el tiempo, tal como la difusión de calor en un medio, en general, son problemas de propagación transitorios sujetos a condiciones iniciales y de frontera. la EDP parabólicas tienen las siguiente forma:

$$\partial_t u - k\partial^2_{xx} u = f \tag{7.29}$$

con las siguientes condiciones iniciales y de frontera:

1. Condición inicial:

$$u(x,0) = f(x) \tag{7.30}$$

2. Condiciones de frontera:

$$\begin{aligned} u(a,t) &= h_1(t) \\ u(b,t) &= h_2(t) \end{aligned} \tag{7.31}$$

Si las condiciones de frontera no dependen del tiempo, el sistema evolucionará hasta un estado de equilibrio donde $\partial_t u = 0$. En el estado estacionario se anula este término de la ecuación, de modo que se puede trabajar la ecuación diferencial como elíptica.

Ecuación de Calor:

$$\frac{\partial u(x,t)}{\partial t} = k\frac{\partial^2 u(x,t)}{\partial x^2} \tag{7.32}$$

donde la región de solución es: $x \in \mathbb{R} = [a,b],\ t \geq 0$. La ecuación diferencial tiene las siguientes condiciones iniciales y de frontera:

1. Condición inicial:

$$u(x,0) = f(x) \tag{7.33}$$

2. Condiciones de frontera:

$$\begin{aligned} u(a,t) &= h_1(t) \\ u(b,t) &= h_2(t) \end{aligned} \tag{7.34}$$

Discretizando el dominio tenemos la siguiente notación: $u(x_i,t_l) = u_i^l$, donde $i = 1,2,...,nx-1$ y $l = 1,2,...,nt-1$. En diferencias finitas, los operadores diferenciales tienen la siguiente forma:

$$\frac{f(x_i,t_{l+1}) - f(x_i,t_l)}{\Delta t} = k(\frac{f(x_{i+1},t_l) - 2f(x_i,t_l) + f(x_{i-1},t_l)}{(\Delta x)^2}) \tag{7.35}$$

para el campo de temperatura tenemos:

$$u_i^{l+1} = u_i^l + \lambda(u_{i+1}^l - 2u_i^l + u_{i-1}^l) \tag{7.36}$$

donde $\lambda = \frac{k\Delta t}{(\Delta x)^2} \leq 1/2$ controla la convergencia del método.

Ejemplo numérico: Calcular las temperaturas al interior de una barra metálica cuya conductividad térmica es $k = 0{,}075\ cm/s$ y longitud $L = 2\ cm$. La fronteras tienen temperaturas $x_0 = 100°$ y $x_f = 50°$ y no dependen del tiempo. El paso temporal es $\Delta t = 0{,}05\ s$ y $\Delta x = 0{,}5\ cm$. Estimar $T_i^{t\to\infty}$.

Para este problema $\lambda = \frac{0{,}075\times 0{,}05}{(0{,}5)^2} = 0{,}015$, lo cual indica que la solución es convergente. Algunos puntos de temperatura son:

x	0	0.5	1	1.5	2.0
T_i^0	100	0	0	0	50
$T_i^{0,05}$	100	1.5	0	0.75	50
$T_i^{0,1}$	100	2.95	0.033	0.47	50
...					
$T_i^{t\to\infty}$	100	87.5	75	62.5	50

Cuadro 7.1: Distribución de temperaturas en la barra para algunos valores temporales, incluyendo la temperatura límite cuando $t \to \infty$.

Ecuación de Calor 2D:

El problema de la ecuación de calor en 2D, se puede definir como sigue:

$$u_t(x,y,t) = \alpha u_{xx}(x,y,t) + \beta u_{yy}(x,y,t) \tag{7.37}$$

donde la región de solución es: $(x,y) \in \mathbb{R} = [a,b] \times [c,d],\ t \geq 0$. La ecuación diferencial tiene las siguientes condiciones iniciales y de frontera:

1. Condición inicial:
$$u(x,y,0) = f(x,y) \tag{7.38}$$

2. Condiciones de frontera:
$$\begin{aligned} u(a,y,t) &= h_1(y,t) \\ u(b,y,t) &= h_2(y,t) \\ u(x,c,t) &= h_3(x,t) \\ u(x,d,t) &= h_4(x,t) \end{aligned} \tag{7.39}$$

La discretización del problema es $u(x_i,x_j,t_l) = u_{i,j}^l$ con $i = 1,2,...,nx-1$, $j = 1,2,...,ny-1$ y $l = 1,2,...,nt-1$. En diferencias finitas, la ecuación diferencial toma la siguiente forma:

$$\begin{aligned} \frac{u(x,y,t+p)-u(x,y,t)}{p} &= \alpha\left(\frac{u(x+h,y,t)-2u(x,y,t)+u(x-h,y,t)}{h^2}\right) \\ &+ \beta\left(\frac{u(x,y+k,t)-2u(x,y,t)+u(x,y-k,t)}{k^2}\right) \end{aligned} \tag{7.40}$$

Definiendo las siguientes constantes:

$$\begin{aligned} \lambda &= \frac{\alpha p}{h^2} \\ \mu &= \frac{\beta p}{k^2} \end{aligned} \tag{7.41}$$

Es posible encontrar un método explícito de la función $u_{i,j,l}$.

$$u_{i,j}^{l+1} = (1-2\lambda-2\mu)u_{i,j}^l + \lambda(u_{i+1,j}^l + u_{i-1,j}^l) + \mu(u_{i,j+1}^l + u_{i,j-1}^l) \tag{7.42}$$

7.3. Problemas hiperbólicos

Los problemas de tipo hiperbólico se refieren a fenómenos que cambian rápido en el tiempo. La ecuación de onda es el ejemplo más inmediato:

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2} \tag{7.43}$$

donde la región de solución es: $x \in \mathbb{R} = [a, b],\ t \geq 0$. La ecuación diferencial tiene las siguientes condiciones iniciales y de frontera:

1. Condición inicial:
$$u(x, 0) = f(x) \tag{7.44}$$

2. Condiciones de frontera:
$$\begin{aligned} u(a, t) &= h_1(t) \\ u(b, t) &= h_2(t) \end{aligned} \tag{7.45}$$

Discretizando el dominio tenemos la siguiente notación: $u(x_i, t_l) = u_i^l$, donde $i = 1, 2, ..., nx - 1$ y $l = 1, 2, ..., nt - 1$. En diferencias finitas, los operadores diferenciales tienen la siguiente forma:

$$\frac{f(x_i, t_{l+1}) - 2f(x_i, t_l) + f(x_i, t_{l-1})}{(\Delta x)^2} = \alpha^2 \left(\frac{f(x_{i+1}, t_l) - 2f(x_i, t_l) + f(x_{i-1}, t_l)}{(\Delta x)^2}\right) \tag{7.46}$$

Para el campo $u(x_i, t_l)$ tenemos el siguiente método explícito.

$$u_i^{l+1} = 2(1 - \lambda^2)u_i^l + \lambda^2(u_{i+1}^l + u_{i-1}^l) - u_i^{l-1} \tag{7.47}$$

donde $\lambda = \frac{\alpha \Delta t}{\Delta x}$. Note que el iterador temporal debe iniciar desde $l = 2$. Generalmente, se inicializa la primera iteración con una condición de tipo Von Neummann.

$$u_i^2 = u_i^1 + \Delta t \left(\frac{\partial u}{\partial t}\right)|_{t=0} \tag{7.48}$$

Para la implementación en `Python` se utiliza la siguiente formula recursiva:

$$u_i^l = 2(1 - \lambda^2)u_i^{l-1} + \lambda^2(u_{i+1}^{l-1} + u_{i-1}^{l-1}) - u_i^{l-2} \tag{7.49}$$

Ecuación de onda 2D:

El problema de la ecuación de onda en 2D, se puede definir como sigue:

$$u_{tt}(x, y, t) = \alpha^2 u_{xx}(x, y, t) + \beta^2 u_{yy}(x, y, t) \tag{7.50}$$

donde la región de solución es: $(x, y) \in \mathbb{R} = [a, b] \times [c, d],\ t \geq 0$. La ecuación diferencial tiene las siguientes condiciones iniciales y de frontera:

1. Condiciones iniciales:

$$\begin{aligned} u(x,y,0) &= f(x,y) \\ u_t(x,y,0) &= g(x,y) \end{aligned} \tag{7.51}$$

2. Condiciones de frontera:

$$\begin{aligned} u(a,y,t) &= h_1(y,t) \\ u(b,y,t) &= h_2(y,t) \\ u(x,c,t) &= h_3(x,t) \\ u(x,d,t) &= h_4(x,t) \end{aligned} \tag{7.52}$$

La discretización del problema es $u(x_i,x_j,t_l) = u^l_{i,j}$ con $i = 1,2,...,nx-1$, $j = 1,2,...,ny-1$ y $l = 1,2,...,nt-1$. En diferencias finitas, la ecuación diferencial toma la siguiente forma:

$$\frac{u^{l+1}_{i,j} - 2u^l_{i,j} + u^{l-1}_{i,j}}{(\Delta t)^2} = \alpha^2(\frac{u^l_{i+1j} - 2u^l_{i,j} + u^l_{i-1j}}{(\Delta x)^2}) + \beta^2(\frac{u^l_{i,j+1} - 2u^l_{i,j} + u^l_{i,j-1}}{(\Delta y)^2}) \tag{7.53}$$

Definiendo $\lambda := \frac{\alpha\Delta t}{\Delta x}$ y $\mu := \frac{\beta\Delta t}{\Delta y}$ tenemos la siguiente expresión explícita:

$$u^{l+1}_{i,j} = 2(1-\lambda^2-\mu^2)u^l_{i,j} + \lambda^2(u^l_{i+1,j} + u^l_{i-1,j}) + \mu^2(u^l_{i,j+1} + u^l_{i,j-1}) + u^{l-1}_{i,j} \tag{7.54}$$

En caso de tener condiciones iniciales de Neummann usamos la diferencia progresiva para hallar $u^1_{i,j}$

$$\begin{aligned} u^1_{i,j} &= u^0_{i,j} + (\Delta t)g(x_i,y_i) \\ &= f(x_i,y_i) + (\Delta t)g(x_i,y_i) \end{aligned} \tag{7.55}$$

Sin embargo, esta opción tiene error de orden $\mathcal{O}(\Delta t)$. Para obtener una solución acorde a los operadores diferenciales de segundo orden se recomienda lo siguiente:

$$\begin{aligned} u(x,y,0+\Delta t) &\cong u(x,y,0) + u_t(x,y,0)\Delta t + \frac{1}{2}u_{tt}(x,y,0)(\Delta t)^2 \\ &= f(x,y) + (\Delta t)g(x,y) + \frac{1}{2}(\Delta t)^2(\alpha^2 u_{xx}(x,y,0) + \beta^2 u_{yy}(x,y,0)) \end{aligned} \tag{7.56}$$

Usando las definiciones de λ y μ se tiene finalmente:

$$\begin{aligned} u^1_{i,j} &= (1-\lambda^2)f(x_i,y_i) + (1-\mu^2)g(x_i,y_i) \\ &+ \frac{\lambda^2}{2}(f(x_{i+1},y_j) + f(x_{i-1},y_j)) + \frac{\mu^2}{2}(f(x_i,y_{j+1}) + f(x_i,y_{j-1})) \end{aligned} \tag{7.57}$$

Ecuación de onda 2D en coordenadas cilíndricas:

Utilizando el operador Laplaciano descrito en la sección anterior es posible escribir un método explícito de solución para la ecuación de onda en coordenadas cilíndricas (Ejercicio). Definamos $\lambda := \frac{\Delta\rho}{\Delta\phi}$ y $\nu := \frac{\alpha\Delta t}{\Delta\rho}$.

$$\begin{aligned} u_{i,j}^{l+1} &= \nu^2\Big[u_{i+1,j}^{l} - 2u_{i,j}^{l} + u_{i-1,j}^{l} \\ &+ \frac{\Delta\rho}{\rho[i]}(u_{i,j}^{l} - u_{i-1,j}^{l}) \\ &+ \left(\frac{\lambda}{\rho[i]}\right)^2(u_{i,j+1}^{l} - 2u_{i,j}^{l} + u_{i,j-1}^{l})\Big] \\ &+ 2u_{i,j}^{l} - u_{i,j}^{l-1} \end{aligned} \tag{7.58}$$

Note que esta implementación del operador de Laplace resulta diferente que el mostrado en la sección de EDP elípticas, dado que se requiere un forma explícita para el tiempo, no para la posición.

7.4. Ejercicios

1. Resolver la ecuación diferencial de conducción de calor 2D en el siguiente dominio rectangular: $(x, y) \in \mathbb{R} = [0, 1] \times [0, 1],\ t \geq 0$. $\alpha = \beta = 1$, la discretización es $\Delta x = \Delta y = 0{,}2$ y $\Delta t = 0{,}1$. $T_{max} = 1\ s$. La ecuación diferencial tiene las siguientes condiciones:

   a) Condición inicial:

   $$u(x, y, 0) = sin(\pi(x + y)) \tag{7.59}$$

   b) Condiciones de frontera:

   $$\begin{aligned} u(0, y, t) &= e^{-2\pi^2 t} sin(\pi y) \\ u(0, y, t) &= e^{-2\pi^2 t} sin(\pi x) \\ u(x, 1, t) &= e^{-2\pi^2 t} sin(\pi(1 + y)) \\ u(x, 1, t) &= e^{-2\pi^2 t} sin(\pi(1 + x)) \end{aligned} \tag{7.60}$$

2. Demuestre la Ecuación (7.58).

Capítulo 8

The finite element method

The finite element method is introduced for solving partial differential equations. It is a commonly used approach and has many advantages. One of them is that it is generally applicable to many problems. In others word, the framework is general enough and applicable to different geometries and materials. We can summarize this method as follows:

1. Discretize the domain.
2. Derive the simpler finite-element equations, which conceptually results difficult to obtain.
3. Assemble (combine) the element equations in a global matrices.
4. Apply the boundary constrains or the initial conditions.
5. Solving the system of linear (non-linear) equations.

In this method, the discretization of the domain is similar to the finite difference method. Nevertheless, the building of the element matrix is completely different. In a 1D lattice, we can think that the basic element is a spring of constant k and length L. The displacement function can be represented as a first-order polynomial.

$$u(x) = a_0 + a_1 x, \tag{8.1}$$

either side of the spring is called *node*. The left-side node is characterized by the displacement u_0 and the right-side node has a displacement u_1. By applying the boundary conditions at $x = 0$ and $x = L$, we get:

$$\begin{aligned} u_0 &= a_0 \\ u_1 &= a_0 + a_1 L \end{aligned} \tag{8.2}$$

The displacement function can be written as:

$$u(x) = (1 - x/L)u_0 + (x/L)u_1 \tag{8.3}$$

Note that, 1 element has 2 degrees of freedom ($[u_0, u_1]$). In general, N elements has $N + 1$ degrees of freedom. The displacement function can also be written in matrix form:

$$u(x) = \begin{pmatrix} 1 - x/L & x/L \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} = \phi^T(x) \cdot \vec{u}(t) \tag{8.4}$$

The vector $\vec{u}(t)$ can be a time-dependent quantity and the ϕ functions are called shape functions, which codify the spatial dependency of the displacement function. The shape-function in each node is equal to 1 and is 0 in the other nodes. In addition, the shape functions satisfy:

$$\sum_{i=0}^{1} \phi_i = 1 \tag{8.5}$$

Using the Hooke's Law ($F = k\delta$), we can calculate the force in each node:

$$\begin{aligned} F_0 &= k(u_0 - u_1) \\ F_1 &= k(u_1 - u_0) \end{aligned} \tag{8.6}$$

in matrix form, we get:

$$\vec{F} = \mathbb{K}\vec{u} \tag{8.7}$$

where $\mathbb{K}$ is known as *stiffness matrix* of the element.

$$\mathbb{K} = \begin{pmatrix} k & -k \\ -k & k \end{pmatrix} \tag{8.8}$$

8.1. Stiffness matrix 2D

If the element is located in the plane and it supports shear and axial stresses the stiffness matrix calculation requires a coordinates transformation. The local reference frame is tagged as S' and the global is tagged as S. The transformation rule is:

$$\begin{pmatrix} u'_0 \\ u'_1 \end{pmatrix} = \begin{pmatrix} cos(\theta) & sin(\theta) \\ -sin(\theta) & cos(\theta) \end{pmatrix} \begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \tag{8.9}$$

In addition, the shear and axial stresses are represented by: F_y and F_x respectively. The extended matrix (by construction) is given by:

$$\begin{pmatrix} f'_{1x} \\ f'_{2x} \\ f'_{1y} \\ f'_{2y} \end{pmatrix} = k \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u'_{1x} \\ u'_{2x} \\ u'_{1y} \\ u'_{2y} \end{pmatrix} \tag{8.10}$$

In the continuum mechanics $k = \frac{AE}{L}$, where E is the Young's modulus and A and L are the area and length respectively. To build the extended matrix of the transformation can be represented through the direct product:

$$\mathbb{I} \otimes \mathbb{T} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} cos(\theta) & sin(\theta) \\ -sin(\theta) & cos(\theta) \end{pmatrix} \tag{8.11}$$

$$\mathbb{T}_{4\times4} = \begin{pmatrix} cos(\theta) & sin(\theta) & 0 & 0 \\ -sin(\theta) & cos(\theta) & 0 & 0 \\ 0 & 0 & cos(\theta) & sin(\theta) \\ 0 & 0 & -sin(\theta) & cos(\theta) \end{pmatrix} \tag{8.12}$$

The transformation of the stiffness matrix is finally given by:

$$\begin{aligned} \vec{F'} &= \mathbb{K}'\vec{u'} \\ \mathbb{T}\vec{F} &= \mathbb{K}'\mathbb{T}\vec{u} \\ \vec{F} &= \mathbb{T}^T\mathbb{K}'\mathbb{T}\vec{u} \end{aligned} \tag{8.13}$$

Explicitly:

$$\mathbb{K} = k \begin{pmatrix} cos^2(\theta) & cos(\theta)sin(\theta) & -cos^2(\theta) & -cos(\theta)sin(\theta) \\ cos(\theta)sin(\theta) & sin^2(\theta) & -cos(\theta)sin(\theta) & -sin^2(\theta) \\ -cos^2(\theta) & -cos(\theta)sin(\theta) & cos^2(\theta) & cos(\theta)sin(\theta) \\ -cos(\theta)sin(\theta) & -sin^2(\theta) & cos(\theta)sin(\theta) & sin^2(\theta) \end{pmatrix} \tag{8.14}$$

8.2. Assembly of elements

The stiffness matrix is a *local* object, which is related to the element itself. The finite element method requires the assembling of a set of elements to obtain the result of a specific problem. The assembled object is a *global* object that contains all elements with their couplings. An example can be the coupling between two objects. Two elements of constants k_1 and k_2 are connected at node 1. The local matrices are given by:

$$\mathbb{K}_1 = \begin{pmatrix} k_1 & -k_1 \\ -k_1 & k_1 \end{pmatrix} \qquad \mathbb{K}_2 = \begin{pmatrix} k_2 & -k_2 \\ -k_2 & k_2 \end{pmatrix} \tag{8.15}$$

Node 2 is the point of coupling of the two elements. In that point, we have the combined effect of those elements. The assembled matrix is given by:

$$\mathbb{K}_{12} = \begin{pmatrix} k_1 & -k_1 & 0 \\ -k_1 & k_1 + k_2 & -k_2 \\ 0 & -k_2 & k_2 \end{pmatrix} \tag{8.16}$$

Suppose that, in the right side a constant force F is applied and the node 0 is fixed ($u_0 = 0$). We can calculate the displacement in each node using the assembled matrix. Hence:

$$\begin{pmatrix} 0 \\ 0 \\ F \end{pmatrix} = \begin{pmatrix} k_1 & -k_1 & 0 \\ -k_1 & k_1 + k_2 & -k_2 \\ 0 & -k_2 & k_2 \end{pmatrix} \begin{pmatrix} 0 \\ u_1 \\ u_2 \end{pmatrix} \tag{8.17}$$

By inverting the assembled matrix, the displacements are given by:

$$\begin{aligned} u_0 &= 0 \\ u_1 &= F/k_1 \\ u_2 &= Fk_1k_2/(k_1+k_2) \end{aligned} \tag{8.18}$$

This result is commonly obtained using static requirements in classical mechanics. Note that, this approach is only the *superposition* of a basic effect through the concept of the global stiffness matrix.

8.3. Dynamical problems

The longitudinal wave function in a beam is described by:

$$u_{tt}(x,t) = c^2 u_{xx}(x,t) \tag{8.19}$$

where $c^2 = E/\rho$. The dynamics of the system can be mapped to a collection of infinity harmonic oscillators:

$$\mathbb{M}\ddot{u}(t) + \mathbb{K}u(t) = \vec{F}(t) \tag{8.20}$$

where $\mathbb{M}$ is the mass matrix of the collection of oscillators and $\vec{F}(t)$ is a generalized force associated to sources. This set must satisfy the initial and boundary conditions. The mass matrix is built using the definition of kinetic energy for continuum media:

$$\begin{aligned} T &= \frac{1}{2}\rho A \int_0^l \dot{u}^2(x,t)dt \\ &= \frac{1}{2}\rho A \int_0^l \dot{\vec{u}}^T \phi(x)\cdot\phi^T(x)\dot{\vec{u}}dx \end{aligned} \tag{8.21}$$

by comparing with the discrete expression of the kinetic energy:

$$T = \frac{1}{2}\dot{\vec{u}}^T\mathbb{M}\dot{\vec{u}} \tag{8.22}$$

$$\begin{aligned} \mathbb{M} &= \rho A \int_0^l \phi(x)\cdot\phi^T(x)dx \\ &= \rho A \int_0^l \begin{pmatrix} \phi_0\phi_0 & \phi_0\phi_1 \\ \phi_1\phi_0 & \phi_1\phi_1 \end{pmatrix} dx \\ &= \frac{\rho A l}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \end{aligned} \tag{8.23}$$

In addition, the stiffness matrix requires the strain energy for continuum media:

$$\begin{aligned} u &= \sigma\epsilon = \frac{EA}{2}\int_0^l u_x^2(x,t)dx \\ &= \frac{1}{2}\vec{u}^T EA\int_0^1 \phi(x)_x\phi(x)_x^T dx\vec{u} \end{aligned} \tag{8.24}$$

by comparing with the discrete expression of the potential energy:

$$u = \frac{1}{2}\vec{u}^T\mathbb{K}\vec{u} \tag{8.25}$$

$$\begin{aligned} \mathbb{K} &= EA\int_0^1 \begin{pmatrix} \phi_{0x}\phi_{0x} & \phi_{0x}\phi_{1x} \\ \phi_{1x}\phi_{0x} & \phi_{x1}\phi_1 x \end{pmatrix} dx \\ &= \frac{EA}{l}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \end{aligned} \tag{8.26}$$

The generalized forces can be calculated:

$$Q_i(t) = \int_0^l f(x,t)\phi_i(x)dx. \tag{8.27}$$

The general expression for n elements is:

$$\begin{aligned} \mathbb{M}_n &= C_m/(6n)\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \\ \mathbb{K}_n &= C_k n\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \end{aligned} \tag{8.28}$$

where $C_m = \rho Al$ and $C_k = EA/l$. Additionally, the following definition is useful:

$$a = \sqrt{C_k/C_m} = \sqrt{E/\rho}/l = c/l \tag{8.29}$$

Finally, the normal modes of oscillation are calculated as follows:

$$det|\mathbb{K} - \omega^2\mathbb{M}| = 0 \tag{8.30}$$

As an example, we can estimate analytically the frequencies for 1 and 2 elements with $u_0 = 0$:

$$C_k - \omega^2 2/6C_m = 0 \tag{8.31}$$

hence $\omega = a\sqrt{3}$.

For two elements, the mass and stiffness matrices are given by:

$$\mathbb{M} = C_m/12\begin{pmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{pmatrix} \tag{8.32}$$

$$\mathbb{K} = 2C_k \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \tag{8.33}$$

By solving the determinant, we have the following frequencies:

$$\omega^2 = \frac{24}{7}(5 \pm 3\sqrt{2})a \tag{8.34}$$

Hence:

$$\begin{aligned} \omega_0 &\approx 1{,}6114a \\ \omega_1 &\approx 5{,}6293a \end{aligned} \tag{8.35}$$

This is an interesting result, which can be compared with the exact value of the wave equation. Note that, the boundary condition in $x = l$:

$$Acos(\omega_n l/c) = 0. \tag{8.36}$$

Using the periodicity condition and the definition of c, we get:

$$\omega_n/a = n\pi/2 \tag{8.37}$$

In other words, using only two elements the fundamental frequency has an error of:

$$\epsilon = \frac{|1{,}6114 - \pi/2|}{\pi/2} \approx 2{,}6\,\% \tag{8.38}$$

References

[1] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927, `https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1927.0118`.

[2] Rubin H Landau, José Páez, Manuel José Páez Mejía, and Cristian C Bordeianu. *A survey of computational physics: introductory computational science*. Princeton University Press, 2008, `https://psrc.aapt.org/items/detail.cfm?ID=11578`.

[3] Ronald E Walpole, Raymond H Myers, Sharon L Myers, and Keying Ye. *Probabilidad y estadística para ingeniería y ciencias*. 4 Edición, 2007.