Mètodes numèrics i probabilístics

Grau en Matemàtica Computacional i Analítica de Dades de la UAB

Carles Barril Basil

Capítol 1

Integració numèrica

Recordatori Interpolació de Lagrange i d'Hermite

Donats $m+1$ punts d'interpolació (x_k, f_k) amb $k = 0 \div m$, existeix un únic polinomi $p_m(x)$ de grau menor o igual a m, anomenat **polinomi interpolador de Lagrange**, tal que

$$p_m(x_k) = f_k \qquad \forall k = 0 \div m.$$

L'expressió de p_m es pot donar amb la fórmula d'interpolació de Lagrange

$$p_m(x) = \sum_{k=0}^{m} f_k l_k(x) \qquad \text{amb} \qquad l_k(x) = \prod_{i=0, i\neq k}^{m} \frac{x - x_i}{x_k - x_i}$$

Si $f \in C^{m+1}$ i p_m és el polinomi interpolador de f sobre els nodes $x_0, x_1, \ldots, x_m$ (és a dir p_m interpola els punts $(x_k, f(x_k))$ amb $k = 0 \div m$), aleshores l'error d'interpolació és

$$f(x) - p_m(x) = \frac{f^{(m+1)}(\xi(x))}{(m+1)!}(x - x_0)(x - x_1)\cdots(x - x_m)$$

on $\xi(x)$ és un punt de l'interval d'interpolació (i.e. l'interval més petit que conté $x, x_0, x_1, \ldots, x_m$).

També existeix un únic polinomi de grau menor o igual a $2m+1$ que, a més a més de coincidir amb una funció f diferenciable sobre els nodes $x_0, x_1, \ldots, x_m$, té la mateixa derivada que f sobre aquests nodes. S'anomena **polinomi interpolador d'Hermite**, i es pot expressar com

$$p_{2m+1}(x) = \sum_{k=0}^{m} f_k(1 - 2l_k'(x_k)(x - x_k))l_k^2(x) + \sum_{k=0}^{m} f_k'(x - x_k)l_k^2(x),$$

amb $f_k = f(x_k)$ i $f_k' = f'(x_k)$. Si $f \in C^{2m+2}$, l'error d'interpolació és

$$f(x) - p_{2m+1}(x) = \frac{f^{(2m+2)}(\xi(x))}{(2m+2)!}(x - x_0)^2(x - x_1)^2\cdots(x - x_m)^2$$

on $\xi(x)$ és un punt de l'interval d'interpolació (i.e. l'interval més petit que conté $x, x_0, x_1, \ldots, x_m$).

1.1 Integració numèrica per quadratures

Donada una funció f definida sobre $[a,b]$ volem calcular

$$J(f) = \int_a^b f(x)dx.$$

La idea que desenvoluparem en aquest tema per calcular numèricament aquesta integral consisteix en discretitzar convenientment el domini $[a,b]$ en una sèrie de nodes $x_0, x_1, \ldots, x_m$ i aproximar la integral mitjançant una suma de la forma

$$J_m(f) = \sum_{k=0}^{m} W_k f(x_k)$$

on W_k s'anomenen pesos d'integració. Aquestes aproximacions es coneixen amb el nom de quadratures ja que en cert sentit el que es fa és aproximar l'àrea sota la corba de la funció f utilitzant una forma poligonal adequada.

El problema bàsic consisteix en trobar el nodes i els pesos de manera que l'error d'integració

$$E_m(f) = J(f) - J_m(f)$$

sigui petit.

1.2 Fórmules d'integració interpolatòries

Siguin $a \leq x_0 < x_1 < x_2 < \cdots < x_m \leq b$ un conjunt de $m+1$ nodes del segment $[a,b]$, i $p_m(x)$ el polinomi interpolador de f sobre aquests nodes. La fórmula interpolatòria sobre aquests $m+1$ nodes consisteix en aproximar la integral $J(f)$ per $J(p_m)$, és a dir:

$$\int_a^b f(x)dx \approx \int_a^b p_m(x)dx.$$

Això equival a dir que

$$\int_a^b f(x)dx \approx \sum_{k=0}^{m} W_k f(x_k) \qquad \text{amb} \qquad W_k = \int_a^b l_k(x)dx = \int_a^b \prod_{i=0, i\neq k}^{m} \frac{x - x_i}{x_k - x_i} dx.$$

Observeu que els pesos W_k depenen de l'interval $[a,b]$ i dels nodes $x_0, x_1, \ldots, x_m$, però no de la funció f.

Exemple. La fórmula dels trapezis és la fórmula interpolatòria que s'obté prenent $m = 1$ i els nodes $x_0 = a$ i $x_1 = b$.

Per la unicitat del polinomi interpolador, **una fórmula interpoladora de $m+1$ nodes és exacte per a qualsevol polinomi de grau menor o igual a m** (perquè el polinomi interpolador sobre

els $m+1$ nodes serà el propi polinomi que pretenem interpolar). Això dóna d'una banda una mesura del grau de precisió de les fórmules interpoladores. Es diu que una fórmula interpoladora és d'**ordre** r si la fórmula és exacte per a tots els polinomis de grau menor o igual a r i falla per algun polinomi de grau $r+1$. Pel que hem comentat, una fórmula interpoladora d'$m+1$ nodes és com a mínim d'ordre m, però veurem que en alguns casos l'ordre és inclús superior. La unicitat del polinomi interpolador també permet calcular els pesos W_k sense necessitat d'integrar els polinomis $l_k(x)$: n'hi ha prou en imposar que l'aproximació sigui certa pels monomis $1, x, x^2, \ldots, x^m$ i resoldre el sistema lineal resultant per a les incògnites W_k per $k = 0 \div m$, que serà un sistema de $m+1$ equacions amb $m+1$ incògnites amb una única solució (aquesta estratègia es coneix com el mètode dels coeficients indeterminats). La proposició següent mostra que si l'aproximació és exacte pels monomis $1, x, x^2, \ldots, x^m$ aleshores també és exacte per tot polinomi de grau més petit o igual a m (és conseqüència del fet que J i J_m són funcions lineals, i.e. del fet que $J(\lambda f+\mu g) = \lambda J(f)+\mu J(g)$ i $J_m(\lambda f + \mu g) = \lambda J_m(f) + \mu J_m(g)$). Després de la proposició es dona un exemple d'aplicació del mètode dels coeficients indeterminats.

Proposició. Si $f = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_m x^m$ i $J_m(x^r) = J(x^r)$ per tot $r \in 0 \div m$, aleshores $J_m(f) = J(f)$.

Prova. En efecte,

$$\begin{aligned}
J_m(f) &= \sum_{k=0}^{m} W_k f(x_k) = \sum_{k=0}^{m} W_k(\alpha_0 + \alpha_1 x_k + \alpha_2 x_k^2 + \cdots + \alpha_m x_k^m) \\
&= \alpha_0 \sum_{k=1} W_k + \alpha_1 \sum_{k=1} W_k x_k + \alpha_2 \sum_{k=1} W_k x_k^2 + \cdots + \alpha_m \sum_{k=1} W_k x_k^m \\
&= \alpha_0 J_m(1) + \alpha_1 J_m(x) + \alpha_2 J_m(x^2) + \cdots + \alpha_m J_m(x^m) \\
&= \alpha_0 J(1) + \alpha_1 J(x) + \alpha_2 J(x^2) + \cdots + \alpha_m J(x^m) \\
&= \alpha_0 \int_a^b 1dx + \alpha_1 \int_a^b x dx + \alpha_2 \int_a^b x^2 dx + \cdots + \alpha_m \int_a^b x^m dx \\
&= \int_a^b \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_m x^m dx = \int_a^b f(x)dx = J(f)
\end{aligned}$$

□

Exemple. Fórmula de Simpson. Es volen trobar els pesos d'integració w_{-1}, w_0 i w_1 tals que la fórmula

$$\int_{-1}^{1} g(t)dt \approx w_{-1}g(-1) + w_0 g(0) + w_1 g(1)$$

sigui exacte per a tots els polinomis de grau menor o igual a 2.

Imposem l'exactitud per $g(t) = 1$, $g(t) = t$ i $g(t) = t^2$:

$$\begin{cases} 2 = w_{-1} + w_0 + w_1 \\ 0 = -w_{-1} + w_1 \\ \frac{2}{3} = w_{-1} + w_1 \end{cases} \quad \Rightarrow \quad w_1 = w_{-1} = \frac{1}{3},\ w_0 = \frac{4}{3}.$$

És a dir hem obtingut la fórmula

$$\int_{-1}^{1} g(t)dt \approx \frac{1}{3}\left(g(-1) + 4g(0) + g(1)\right).$$

Aquesta fórmula es pot traslladar a qualsevol interval $[a, b]$ utilitzant el canvi de variable

$$x(t) = \frac{b-a}{2}t + \frac{b+a}{2}.$$

En efecte:

$$\begin{aligned}\int_a^b f(x)dx = \int_{-1}^{1} f(x(t))x'(t)dt &= \int_{-1}^{1} f(x(t))\frac{b-a}{2}dt \\ &\approx \frac{b-a}{2}\frac{1}{3}\left(f(x(-1)) + 4f(x(0)) + f(x(1))\right) \\ &= \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)\end{aligned}$$

Observació. El mètode de Simpson és exacte també per polinomis de grau 3:

$$\int_{-1}^{1} t^3 dt = 0 \qquad \text{i} \qquad \frac{1}{3}((-1)^3 + 0^3 + 1^3) = 0,$$

és a dir la integral de t^3 coincideix amb la fórmula aplicada a t^3. La fórmula traslladada a l'interval $[a, b]$ també és exacte per a tots els polinomis de grau menor o igual a 3. En efecte, si $k \leq 3$, es té

$$\int_a^b x^k dx = \frac{b-a}{2}\int_{-1}^{1}\left(\frac{b-a}{2}t + \frac{b+a}{2}\right)^k dt$$

i la integral de la dreta és exacte quan s'utilitza la fórmula de Simpson a l'interval $[-1, 1]$ ja que l'integrand és un polinomi de grau menor o igual a 3, i la fórmula és exacte per aquests polinomis (per construcció fins a polinomis de grau 2 i de "propina", a causa d'una certa simetria de fet, per a polinomis de fins a grau 3 com hem comprovat al principi d'aquesta observació).

1.3 Fórmules de Newton-Cotes

La fórmula dels trapezis i la de Simpson són casos particulars del que es coneix com a fórmules tancades de Newton-Cotes. Aquestes fórmules s'obtenen quan es prenen $m+1$ nodes equiespaiats a l'interval $[a, b]$:

$$x_k = a + kh, \qquad k = 0 \div m \qquad \text{i} \qquad h = \frac{b-a}{m},$$

i aleshores

$$\int_a^b f(x)dx \approx \sum_{k=0}^{m} W_k f(x_k).$$

Recordem que

$$W_k = \int_a^b \prod_{i=0, i\neq k}^{m} \frac{x - x_i}{x_k - x_i} dx,$$

de manera que, com $x_i = a + ih$, W_k es pot expressar com

$$\begin{aligned} W_k &= \int_a^b \prod_{i=0, i\neq k}^{m} \frac{x - a - ih}{a + kh - a - ih} dx \\ &= \int_a^b \prod_{i=0, i\neq k}^{m} \frac{\frac{x-a}{h} - i}{k - i} dx \\ &= h \int_0^m \prod_{i=0, i\neq k}^{m} \frac{t - i}{k - i} dt = h\alpha_k \end{aligned}$$

on a la penúltima igualtat s'ha utilitzat el canvi de variable $t = (x - a)/h$. Observem que els coeficients α_k no depenen ni de l'interval $[a, b]$ ni de la funció f. En termes dels coeficients α_k la fórmula s'escriu

$$\boxed{\mathrm{NC}_m(f, [a, b]) := h \sum_{k=0}^{m} \alpha_k f(x_k) \qquad \text{amb} \qquad h = \frac{b - a}{m} \qquad \text{i} \qquad x_k = a + kh.}$$

Proposició. L'error de la fórmula de Newton-Cotes satisfà:

$$E_m(f, [a, b]) := \int_a^b f(x) - p_m(x) dx = \int_a^b f(x) dx - \mathrm{NC}_m(f, [a, b]) = K_m \frac{f^{(p+1)}(\xi)}{(p + 1)!} h^{p+2}$$

amb $\xi \in (a, b)$ i

- Si m és senar, $p = m$ i $K_m = E_m(x^{m+1}, [a, b])/h^{m+2}$
- Si m és parell, $p = m + 1$ i $K_m = E_m(x^{m+2}, [a, b])/h^{m+3}$

Exemple. Per trobar el factor K_2 pel cas $m = 2$ (és a dir per la fórmula de Simpson) observem que $h = (b - a)/2$ i calculem

$$\begin{aligned} K_2 = E_2(x^4, [a, b])/h^5 &= \frac{2^5}{(b - a)^5} \left(\int_a^b x^4 dx - \frac{b - a}{6} \left(a^4 + 4 \left(\frac{a + b}{2} \right)^4 + b^4 \right) \right) \\ &= \frac{2^5}{(b - a)^5} \left(\frac{1}{5}(b^5 - a^5) - \frac{b - a}{6} \left(a^4 + 4 \left(\frac{a + b}{2} \right)^4 + b^4 \right) \right) = -\frac{4}{15} \end{aligned}$$

Comentari. Es pot veure (tot i que no és fàcil de provar) que les constants K_m no depenen de l'interval $[a, b]$.

A continuació es dona una taula amb els coeficients α_k per a les fórmules de Newton-Cotes en funció del nombre de nodes:

m	d	$d \cdot \alpha_k$	K_m	Nom
1	2	1 1	$-1/6$	Trapezis
2	3	1 4 1	$-4/15$	Simpson
3	8/3	1 3 3 1	$-9/10$	3/8
4	45/2	7 32 12 32 7	$-128/21$	Milne

Utilitzar el mètode de Newton-Cotes amb un nombre m elevat (i.e. utiltzar $m+1$ nodes equiespaiats) no és recomenable perquè hi ha funcions que no es poden aproximar arbitràriament bé utilitzant polinomis interpoladors sobre nodes equiespaiats (és el que es coneix com a fenomen de Runge, i l'exemple paradigmàtic és la funció $f(x) = 1/(1+25x^2)$ sobre l'interval $[-1,1]$). Si el polinomi interpolador no aproxima bé la funció, no és esperable que la integral del polinomi interpolador sigui una bona aproximació de la integral de la funció. Una estratègia per obtenir fórmules més precises utilitzant mètodes de Newton-Cotes d'ordre baix consisteix en dividir l'interval d'integració en sub-intervals més petits i aplicar una fórmula interpolatòria d'ordre baix (per exemple Trapezis o Simpson) en cada un dels subintervals. Aquesta estratègia dona lloc a les fórmules compostes.

A continuació es dona la prova de la proposició sobre l'error de les fórmules de Newton-Cotes pel cas $m = 1$. Per poder donar la demostració és oportú introduir primer el següent resultat.

Teorema del valor mitja per integrals. Sigui $w : [a,b] \to [0,\infty)$ una funció integrable no negativa. Aleshores per tota $f : [a,b] \to \mathbb{R}$ continua es té

$$\int_a^b f(x)w(x)dx = f(\xi)\int_a^b w(x)dx \qquad \text{per algun } \xi \text{ de } (a,b).$$

Prova. En efecte, per ser f conínua en $[a,b]$ exiteixen $x_{\min}, x_{\max} \in [a,b]$ tals que per tot $x \in [a,b]$ es té

$$f(x_{\min}) \le f(x) \le f(x_{\max}).$$

Ara, com que w és no negativa, es té

$$f(x_{\min})\int_a^b w(x)dx \le \int_a^b f(x)w(x)dx \le f(x_{\max})\int_a^b w(x)dx, \tag{1.1}$$

i això implica, per la continuïtat de f, que existexi $\xi \in I(x_{\min}, x_{\max}) \subset (a,b)$, on $I(x_{\min}, x_{\max})$ denota l'interval obert amb extrems $x_{\min}$ i $x_{\max}$, tal que

$$f(\xi)\int_a^b w(x)dx = \int_a^b f(x)w(x)dx.$$

Per veure això més clarament podeu reescriure (1.1) com

$$f(x_{\min})\int_a^b w(x)dx - \int_a^b f(x)w(x)dx \le 0 \le f(x_{\max})\int_a^b w(x)dx - \int_a^b f(x)w(x)dx$$

i aplicar el Teorema de Bolzano a la funció $\xi \mapsto f(\xi)\int_a^b w(x)dx - \int_a^b f(x)w(x)dx$ definida a l'interval $I(x_{\min}, x_{\max})$, la qual cosa assegura l'existencia d'un valor $\xi \in I(x_{\min}, x_{\max})$ tal que

$$f(\xi)\int_a^b w(x)dx - \int_a^b f(x)w(x)dx = 0. \qquad \square$$

Prova de la proposició per $m = 1$ (Trapezis). Recordem que si $p_1(x)$ és el polinomi interpolador de la funció f sobre els nodes $x_0 = a$ i $x_1 = b$, aleshores es té

$$f(x) = p_1(x) + \frac{f''(\xi(x))}{2}(x-a)(x-b),$$

de manera que

$$E_1(f,[a,b]) := \int_a^b f(x) - p_1(x)dx = \int_a^b \frac{f''(\xi(x))}{2}(x-a)(x-b)dx.$$

Observem ara que la funció $w(x) = -(x-a)(x-b)$ es no negativa a $[a,b]$, per tant pel Teorema del valor mitja per integrals es té

$$\int_a^b \frac{f''(\xi(x))}{2}(x-a)(x-b)dx = -\frac{f''(\xi(\eta))}{2}\int_a^b w(x)dx = \frac{f''(\xi(\eta))}{2}\int_a^b (x-a)(x-b)dx$$

on $\xi(\eta) \in (a,b)$ perquè $\eta \in (a,b)$, i com que η està fixat, podem escriure ξ enlloc de $\xi(\eta)$, és a dir es té

$$E_1(f,[a,b]) = \frac{f''(\xi)}{2}\int_a^b (x-a)(x-b)dx,$$

com es volia veure. De propina s'obté una expressió alternativa per K_1, només cal aplicar la última fórmula a $f(x) = x^2$ de manera que

$$K_1 := \frac{E_1(x^2,[a,b])}{(b-a)^3} = \frac{\int_a^b (x-a)(x-b)dx}{(b-a)^3}.$$

$\square$

1.4 Fórmules compostes

Per calcular la integral

$$\int_a^b f(x)dx$$

emprant una fórmula de Newton-Cotes de $m+1$ punts composta cal:

- Dividir l'interval $[a,b]$ en I subintervals:

$$[x_i, x_{i+1}] \quad \text{on} \quad x_i = a + i\frac{b-a}{I} \quad \text{amb} \quad i \in 0 \div I.$$

- Aplicar la fórmula de Newton-Cotes de $m+1$ punts a cada subinterval i sumar els resultats. Concretament s'escriu

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \cdots + \int_{x_{m-1}}^{x_m} f(x)dx$$

i s'utilitza

$$\int_{x_i}^{x_{i+1}} f(x)dx = \mathrm{NC}_m(f,[x_i,x_{i+1}]) + E_m(f,[x_i,x_{i+1}]) = \mathrm{NC}_m(f,[x_i,x_{i+1}]) + K_m \frac{f^{(p+1)}(\xi_i)}{(p+1)!}\left(\frac{b-a}{mI}\right)^{p+2}$$

on $\xi_i \in (x_i, x_{i+1})$ i $p = m$ si m és senar i $p = m+1$ si m és parell. Així es té

$$\int_a^b f(x)dx = \sum_{i=0}^{I-1} \mathrm{NC}_m(f,[x_i,x_{i+1}]) + K_m \frac{1}{(p+1)!}\left(\frac{b-a}{mI}\right)^{p+2} \sum_{i=0}^{I-1} f^{(p+1)}(\xi_i).$$

Si $f^{(p+1)}$ és contínua, el terme d'error de la fórmula composta

$$\mathrm{EC}_m(f,[a,b]) := \frac{K_m}{(p+1)!}\left(\frac{b-a}{mI}\right)^{p+2} \sum_{i=0}^{I-1} f^{(p+1)}(\xi_i)$$

es pot escriure com

$$\mathrm{EC}_m(f,[a,b]) = \frac{K_m}{(p+1)!}\frac{b-a}{m}\left(\frac{b-a}{mI}\right)^{p+1} f^{(p+1)}(\xi) \tag{1.2}$$

amb $\xi \in (a,b)$, la qual cosa facilita l'anàlisi de l'error en casos pràctics. Per veure aquesta igualtat observem que si $f \in C^{p+1}(a,b)$, aleshores

$$\frac{1}{I}\sum_{i=0}^{I-1} f^{(p+1)}(\xi_i) = f^{(p+1)}(\xi) \tag{1.3}$$

per algun $\xi \in (a,b)$. En efecte, com que la part esquerra de la igualtat anterior és la mitjana aritmètica dels valors $\{f^{(p+1)}(\xi_0), f^{(p+1)}(\xi_1), \ldots, f^{(p+1)}(\xi_{I-1})\}$, necessàriament existeix un valor en aquest conjunt que és menor que la mitjana i un valor en aquest conjunt que es major que la mitjana. En altres paraules, existeixen valors $\xi_{\min}, \xi_{\max} \in \{\xi_0, \xi_1, \xi_2, \ldots, \xi_{m-1}\}$ tals que

$$f^{(p+1)}(\xi_{\min}) \le \frac{1}{I}\sum_{i=0}^{I-1} f^{(p+1)}(\xi_i) \le f^{(p+1)}(\xi_{\max}). \tag{1.4}$$

Ara bé, com que la funció $f^{(p+1)}$ és contínua, existeix un valor ξ contingut a l'interval obert que té per extrems els valors $\xi_{\min}$ i $\xi_{\max}$ tal que

$$f^{(p+1)}(\xi) = \frac{1}{I}\sum_{i=0}^{I-1} f^{(p+1)}(\xi_i).$$

Per veure més clarament això podeu reescriure (1.4) com

$$f^{(p+1)}(\xi_{\min}) - \frac{1}{I}\sum_{i=0}^{I-1} f^{(p+1)}(\xi_i) \le 0 \le f^{(p+1)}(\xi_{\max}) - \frac{1}{I}\sum_{i=0}^{I-1} f^{(p+1)}(\xi_i)$$

i aplicar el Teorema de Bolzano per concloure que existeix ξ entre $\xi_{\min}$ i $\xi_{\max}$ tal que

$$f^{(p+1)}(\xi) - \frac{1}{I}\sum_{i=0}^{I-1} f^{(p+1)}(\xi_i) = 0.$$

Exercici. En quants subintervals s'ha de dividir l'interval $[0,1]$ per calcular la integral

$$\int_0^1 e^{-x^2} dx$$

amb un error menor a ε si s'utilitza la fórmula dels Trapezis composta (donar el nombre d'intervals en funció de valor ε).

Sabem que l'error de la fórmula composta és

$$E = \mathrm{EC}_1(f,[0,1]) = -\frac{1}{12}\frac{1}{I^2} f''(\xi)$$

amb $f(x) = e^{-x^2}$ i $\xi \in (0,1)$, on I és el nombre de subintervals utilitzats. Com que

$$|E| = \frac{1}{12} f''(\xi)\frac{1}{I^2} \leq \frac{1}{12}\max_{\xi\in[0,1]} f''(\xi)\frac{1}{I^2},$$

n'hi ha prou en buscar el valor $I \in \mathbb{N}$ més petit que satisfà

$$\frac{1}{12}\max_{\xi\in[0,1]} f''(\xi)\frac{1}{I^2} < \varepsilon,$$

és a dir el valor I tal que

$$I > \sqrt{\frac{1}{12\varepsilon}\max_{\xi\in[0,1]} f''(\xi)}.$$

Es pot veure (fent la gràfica de $f'' : [0,1] \to \mathbb{R}$ si cal) que

$$\max_{\xi\in[0,1]} f''(\xi) = 2,$$

de manera que la desigualtat anterior esdevé

$$I > \sqrt{\frac{1}{6\varepsilon}}.$$

En particular si $\varepsilon = 10^{-4}$ cal prendre $I = 41 > 40.82 \approx \sqrt{10^4/6}$.

Observeu que aquest procediment per trobar en quants subintervals s'ha de refinar l'interval original $[a,b]$ per tal que l'aproximació de la integral tingui un error menor a ε requereix conèixer cotes de les derivades de la funció f. Això no és molt pràctic si el que es vol és implementar un programa que rebi com argument la funció f i la tolerància ε retorni una aproximació de la integral amb error menor a ε (la feinada que hauria de fer el programa per determinar el màxim i el mínim que assoleix f'' a l'interval d'integració tindria un cost computacional més alt que aproximar la pròpia integral). Més endavant veurem un procediment que permet estimar l'error d'aproximació sense utilitzar informació sobre les derivades de la funció f, la qual cosa permet implementar algorismes d'intergació automàtica. Abans, però, introduïm una altra família de fórmules interpolatòries.

1.5 Fórmules Gaussianes

En les fórmules interpolatòries que hem vist es fixaven els $m+1$ punts d'interpolació i deixàvem com a única llibertat els $m+1$ pesos W_k. Això fa que a priori aquestes fórmules puguin ser a exactes per a polinomis de fins a grau m. A continuació veurem que si es trien els nodes $x_0, x_1, \ldots, x_m$ de forma apropiada, aleshores les fórmules poden ser exactes per polinomis de fins a grau $2m+1$. Per fer-ho cal introduir el concepte de polinomis ortogonals.

Definició. Un conjunt de polinomis $\{\psi_0, \psi_1, \psi_2, \ldots\}$ és una família de polinomis ortogonals respecte un interval $[a,b]$ i una funció pes $w : [a.b] \to (0,\infty)$ (observeu que la funció pes és positiva) si $\text{grau}(\psi_k(x)) = k$ per tot $k \in \mathbb{N} \cup \{0\}$ i

$$\int_a^b w(x)\psi_i(x)\psi_j(x)dx = \begin{cases} 0 & \text{si} \quad i \neq j \\ c_i \neq 0 & \text{si} \quad i = j \end{cases}.$$

Es pot veure que el subconjunt $\{\psi_0(x), \psi_1(x), \psi_2(x), \ldots, \psi_k(x)\}$ és una base de l'espai de polinomis de grau menor o igual a k, és a dir tot polinomi $p(x)$ de grau menor o igual a k s'expressa de forma única com

$$p(x) = \sum_{i=0}^{k} \alpha_i \psi_i(x),$$

en el sentit que existeix una i només una combinació de coeficients α_i que fan certa la igualtat anterior. Per veure-ho n'hi ha prou en igualar els coeficients de p (donats) amb els coeficients del polinomi $\sum_{i=0}^{k} \alpha_i\psi_i(x)$, la qual cosa resulta amb un sistema lineal de $k+1$ equacions amb $k+1$ incògnites (els valors $\alpha_0, \alpha_1, \ldots, \alpha_k$), i gràcies al fet que $\text{grau}(\psi_i(x)) = i$, el sistema és triangular amb elements diferents de zero a la diagonal, la qual cosa implica que el sistema és compatible determinat.

Notació. Per simplificar les expressions, quan sigui oportú denotarem

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

per a tot parell de funcions $f, g : [a,b] \to \mathbb{R}$.

Propitetat. Si $p(x)$ és un polinomi de grau menor a k aleshores

$$\langle p, \psi_k\rangle = 0.$$

Prova. En efecte, $p(x)$ es pot escriure com

$$p(x) = \sum_{i=0}^{k-1} \alpha_i \psi_i(x)$$

de manera que, per la linealitat de la integral, es té

$$\langle p, \psi_k\rangle = \left\langle \sum_{i=0}^{k-1} \alpha_i\psi_i, \psi_k \right\rangle = \sum_{i=0}^{k-1} \alpha_i \langle \psi_i, \psi_k\rangle = 0,$$

on la darrera igualtat és conseqüència de l'ortogonalitat dels polinomis $\psi_i(x)$ i $\psi_k(x)$ quan $k \neq i$.

□

Propietat. El polinomi $\psi_k(x)$ té k zeros simples a l'interval (a,b).

Prova. En efecte, si $\psi_k(x)$ només canviés de signe en els nodes $x_1, x_2, \ldots, x_i \in (a,b)$ amb $i < k$, (la qual cosa implicaria en particular que $\psi_k(x)$ té com a molt $i < k$ zeros simples perquè en tot zero simple d'un polinomi el polinomi presenta necessàriament un canvi de signe), aleshores el polinomi

$$q(x)\psi_k(x)$$

amb $q(x) = (x - x_1)(x - x_2)\ldots(x - x_i)$, no canviaria de signe en tot $[a,b]$ (ja que $q(x)$ i $\psi_k(x)$ canvien de signe en els mateixos punts). Això implicaria que

$$\int_a^b w(x)q(x)\psi_k(x)dx = \langle q, \psi_k \rangle \neq 0,$$

pel fet que la funció pes w és positiva, la qual cosa contradiu la propietat anterior ja que q és de grau $i < k$. Es dedueix, per tant, que $\psi_k(x)$ ha de canviar de signe en almenys k punts continguts a (a,b), que equival a dir que $\psi_k(x)$ té almenys k zeros a (a,b), però $\psi_k(x)$ no pot tenir més de k zeros per ser un polinomi de grau k. Es conclou així que $\psi_k(x)$ té k zeros simples continguts a (a,b).

□

Exemple: polinomis de Legendre. Considerem $[a,b] = [-1,1]$ i la funció pes $w(x) = 1$ sobre aquest interval. Una manera per donar una família de polinomis otrogonals respecte aquest interval i aquesta funció pes és la següent:

- S'agafa un polinomi de grau zero qualsevol, per exemple $\psi_0(x) = 1$.
- Es considera un polinomi de grau 1 genèric, $\psi_1(x) = A_1 x + A_0$ i s'imposa que sigui ortogonal a ψ_0 i no nul, és a dir s'imposa que (tenint en compte que $[a,b] = [-1,1]$ i que $w(x) = 1$)

$$\begin{cases} 0 = \displaystyle\int_{-1}^{1} \psi_0(x)\psi_1(x)dx = \int_{-1}^{1} A_1 x + A_0\, dx \\ 0 \neq \displaystyle\int_{-1}^{1} \psi_1(x)\psi_1(x)dx = \int_{-1}^{1} (A_1 x + A_0)^2\, dx \end{cases}.$$

  Observem que de la primera equació se segueix

$$0 = \left[\frac{A_1 x^2}{2} + A_0 x\right]_{x=-1}^{x=1} = 2A_0 \quad \Rightarrow \quad A_0 = 0,$$

  i utilitzant $A_0 = 0$ a la segona equació es té

$$0 \neq \left[\frac{A_1^2 x^3}{3}\right]_{x=-1}^{x=1} = \frac{2}{3}A_1^2,$$

  que se satisfà si, per exemple, $A_1 = 1$. Per tant podem agafar $\psi_1 = x$.

- Es considera un polinomi de grau 2 genèric, $\psi_2(x) = A_2x^2 + A_1x + A_0$ i s'imposa que sigui ortogonal a ψ_0 i ψ_1 i no nul, és a dir s'imposa que

$$\begin{cases} 0 = \displaystyle\int_{-1}^{1} \psi_0(x)\psi_2(x)dx = \int_{-1}^{1} A_2x^2 + A_1x + A_0\, dx \\ 0 = \displaystyle\int_{-1}^{1} \psi_1(x)\psi_2(x)dx = \int_{-1}^{1} x(A_2x^2 + A_1x + A_0)\, dx\ . \\ 0 \neq \displaystyle\int_{-1}^{1} \psi_2(x)\psi_2(x)dx = \int_{-1}^{1} (A_2x^2 + A_1x + A_0)^2\, dx \end{cases}$$

  De la primera equació es té

$$0 = \left[\frac{A_2x^3}{3} + A_1\frac{x^2}{2} + A_0x\right]_{x=-1}^{x=1} = \frac{2}{3}A_2 + 2A_0 \quad \Rightarrow \quad A_2 = -3A_0,$$

  i de la segona equació

$$0 = \left[\frac{A_2x^4}{4} + A_1\frac{x^3}{3} + A_0\frac{x^2}{2}\right]_{x=-1}^{x=1} = \frac{2}{3}A_1 \quad \Rightarrow \quad A_1 = 0..$$

  Utilitzant aquestes relacions a la tercera equació es té

$$0 \neq \left[\frac{A_2x^5}{5} + 2A_2A_0\frac{x^3}{3} + A_0^2x\right]_{x=-1}^{x=1} = \frac{2}{5}A_2^2 + \frac{4}{3}A_2A_0 + 2A_0^2 = \left(\frac{18}{5} - \frac{12}{3} + 2\right)A_0^2,$$

  que se satisfà si, per exemple, $A_0 = -1$. Per tant podem agafar $\psi_2 = 3x^2 - 1$.

- Etcetera.

D'aquesta manera es genera la família de polinomis ortogonals de Legendre (que és la família de polinomis ortogonals respecte l'interval $[-1, 1]$ i la funció pes $w(x) = 1$).

Observació. La llibertat que hi ha per escollir algun dels coeficients a cada pas fa que hi hagi diferents famílies de polinomis de Legendre. Per evitar aquest fenòmen se sol definir la família de polinomis ortogonals de Legendre "estandaritzada" com la família de polinomis ortogonals respecte l'interval $[-1, 1]$ i la funció pes $w(x) = 1$ que satisfan, a més a més, que $\psi_k(1) = 1$ per tot $k \geq 0$. Noteu que els polinomis $\psi_0(x)$ i $\psi_1(x)$ donats a dalt ja estan estandaritzats, però el polinomi $\psi_2(x)$ no. Perquè ho fos hauríem d'haver agafat $A_0 = -1/2$, ja que aleshores hauríem obtingut $\psi_2(x) = (3x^2 - 1)/2$ que sí satisfà $\psi_2(1) = 1$. Es pot veure que els polinomis de Legendre estandaritzats són únics. En general quan es fa referència als polinomis de Legendre s'entén que s'està parlant de la família estandaritzada.

Fórmula d'integració Gaussiana de Legendre (Fórmula de Gauss-Legendre). És la fórmula interpolatòria de $m + 1$ punts sobre l'interval $[-1, 1]$ que s'obté quan s'utilitzen els zeros del polinomi de Legendre ψ_{m+1} com a nodes d'interpolació.

Fórmula d'integració Gaussiana (General)

(associada a una família de polinomis ortogonals $\{\psi_0, \psi_1, \psi_2, \dots\}$ respecte l'interval $[a, b]$ i la funció pes w)

Considerem la fórmula d'aproximació

$$\int_a^b w(x)f(x)dx \approx \sum_{k=0}^{m} W_k f(x_k)$$

on x_k amb $k = 0 \div m$ són els $m+1$ zeros simples del polinomi ψ_{m+1} en $[a, b]$ (sabem que existeixen per la propietat provada a l'apartat anterior). De la mateixa manera que hem vist amb les fórmules interpolatòries sense pes w, si volem que la fórmula anterior sigui certa per a tots els polinomis de grau menor o igual a m cal imposar que els pesos W_k satisfacin

$$W_k = \int_a^b w(x)l_k(x)dx \qquad \text{on} \qquad l_k(x) = \prod_{i=0, i\neq k}^{m} \frac{x - x_i}{x_k - x_i}, \tag{1.5}$$

ja que si f és un polinomi de grau menor o igual a m, f coincideix amb el polinomi interpolador de grau m, és a dir per aquestes f es té que

$$f(x) = \sum_{k=0}^{m} f(x_k)l_k(x),$$

i imposar que la fórmula sigui exacte en aquests casos implica (1.5).

El cas és que amb aquests pesos (i gràcies a que els nodes són els zeros de $\psi_{m+1}(x)$) la fórmula és exacte per a polinomis de fins a grau $2m + 1$.

Propitetat. Si $p_{2m+1}(x)$ és un polinomi de grau menor o igual a $2m + 1$ aleshores

$$\int_a^b w(x)p_{2m+1}(x)dx = \sum_{k=0}^{m} W_k p_{2m+1}(x_k)$$

Prova. Com que $p_{2m+1}(x)$ és un polinomi de grau menor o igual a $2m+1$, aleshores es pot escriure (dividint el polinomi $p_{2m+1}(x)$ per $\psi_{m+1}(x)$) com

$$p_{2m+1}(x) = q_m(x)\psi_{m+1}(x) + r_m(x)$$

on $q_m(x)$ i $r_m(x)$ tenen com a molt grau m. Llavors, com que

$$\int_a^b w(x)q_m(x)\psi_{m+1}(x)dx = \langle q_m, \psi_{m+1}\rangle = 0$$

per ser $q_m(x)$ de grau menor a $m + 1$, es té

$$\begin{aligned}\int_a^b w(x)p_{2m+1}(x) &= \int_a^b w(x)q_m(x)\psi_{m+1}(x)dx + \int_a^b w(x)r_m(x)dx = \int_a^b w(x)r_m(x)dx \\ &= \sum_{k=0}^{m} W_k r_m(x_k) = \sum_{k=0}^{m} W_k(q_m(x_k)\psi_{m+1}(x_k) + r_m(x_k)) \\ &= \sum_{k=0}^{m} W_k p_{2m+1}(x_k)\end{aligned}$$

on a l'antepenúltima igualtat hem utilitzat que la fórmula interpolatòria és exacte per ser $r_m(x)$ un polinomi de grau menor o igual a m (recordeu que tota fórmula d'integració interpolatòria de $m+1$ punts és exacte per a polinomis de grau m o menor), i a la penúltima igualtat hem utilitzat que x_k són zeros de ψ_{m+1} (de manera que $\psi_{m+1}(x_k) = 0$). És a dir, la fórmula és exacte pel polinomi $p_{2m+1}(x)$.

□

Error de les fórmules Gaussianes

Per a funcions $f \in C^{2m+2}(a,b)$ es pot donar una expressió per a l'error. Per fer-ho considerem el polinomi interpolador d'Hermite (veure el recordatori sobre els polinomis interpoladors) sobre els nodes $x_0, x_1, \ldots, x_m$ (és a dir els zeros de ψ_{m+1}). Per aquest polinomi, que denotem com p_{2m+1}, la fórmula gaussiana (associada a la família de polinomis ortogonals $\{\psi_0, \psi_1, \psi_2, \ldots\}$ respecte l'interval $[a,b]$ i funció pes w) és exacte ja que el grau de p_{2m+1} és $2m+1$ com a molt.

Considerem ara la fórmula per l'error entre $f(x)$ i p_{2m+1}, que és

$$f(x) - p_{2m+1}(x) = \frac{f^{(2m+2)}(\xi(x))}{(2m+2)!}(x-x_0)^2(x-x_1)^2\cdots(x-x_m)^2$$

on $\xi(x)$ és un punt de l'interval d'interpolació (i.e. l'interval més petit que conté $x, x_0, x_1, \ldots, x_m$). Observem que

$$\psi_{m+1}(x) = A_{m+1}(x-x_0)(x-x_1)\cdots(x-x_m)$$

per alguna constant A_{m+1} (que és el coeficient dominant de $\psi_{m+1}(x)$), ja que ambdos polinomis (el de la dreta de la igualtat i el de l'esquerra) tenen el mateix grau i els mateixos zeros. Amb això en ment, ara integrem a banda i banda entre a i b després de multiplicar a banda i banda per la funció pes w. En fer-ho (i tenint en compte que la fórmula Gaussiana és exacte per p_{2m+1}) obtenim

$$\int_a^b w(x)f(x)dx - \sum_{k=0}^{m} W_k f(x_k) = \frac{1}{(2m+2)!}\frac{1}{A_{m+1}^2}\int_a^b f^{(2m+2)}(\xi(x))w(x)\psi_{m+1}^2(x)dx,$$

que es pot reescriure gràcies al Teorema del valor mitjà per integrals (ja que la funció $w(x)\psi_{m+1}^2(x)$ és no negativa en $[a,b]$) com

$$\boxed{\int_a^b w(x)f(x)dx - \sum_{k=0}^{m} W_k f(x_k) = \frac{f^{(2m+2)}(\xi)}{(2m+2)!}\frac{1}{A_{m+1}^2}\int_a^b w(x)\psi_{m+1}^2(x)dx,} \tag{1.6}$$

on $\xi(\eta) \in (a,b)$ perquè $\eta, x_0, x_1, \ldots, x_m \in (a,b)$, de manera que (com η és un valor fixat) es pot escriure $\xi \in (a,b)$ directament enlloc de $\xi(\eta)$ a l'expressió anterior.

Observació. El factor de l'error

$$\frac{1}{A_{m+1}^2}\int_a^b w(x)\psi_{m+1}^2(x)d$$

es pot obtenir aplicant la fórmula Gaussiana a $f(x) = x^{2m+2}$, ja que aleshores

$$\frac{f^{(2m+1)}(\xi)}{(2m+1)!} = 1.$$

Observació (Legendre). Per la fórmula de Gauss-Legendre, l'error es pot simplificar com

$$\boxed{\int_{-1}^{1} f(x) - \sum_{k=0}^{m} W_k f(x_k) = \frac{2^{2m+3}((m+1)!)^4}{(2m+3)((2m+2)!)^3} f^{(2m+2)}(\xi) \qquad \text{amb} \qquad \xi \in (-1,1).}$$

Per més detalls consultar Aubanell-Benseny-Delshams, Problema IV.8 de la pàgina 392.

Càlcul de famílies de polinomis ortogonals: fórmula de recurrència

Vegem un manera més sistemàtica de trobar una família de polinomis ortogonals respecte $[a,b]$ i una funció pes $w(x)$ en general. La idea és trobar una fórmula tancada que ens expressi ψ_{k+1} en termes de $\psi_0, \psi_1, \ldots, \psi_k$ que s'hauran calculat prèviament. Per fer-ho comencem considerant el polinomi $\phi_{k+1} = x\psi_k(x)$. Com que $\{\psi_0, \psi_1, \ldots, \psi_{k+1}\}$ ha de ser una base de l'espai de polinomis de grau menor o igual a $k+1$, existeixen coeficients $\alpha_0, \ldots, \alpha_{k+1}$ tals que

$$\phi_{k+1} = \sum_{l=0}^{k+1} \alpha_l \psi_l(x), \tag{1.7}$$

i a més a més sabem que $\alpha_{k+1} \neq 0$ ja que ϕ_{k+1} té grau $k+1$ de manera que no n'hi ha prou amb els polinomis $\{\psi_0, \ldots, \psi_k\}$ per generar ϕ_{k+1}.

D'altra banda, com que $\{\psi_0, \psi_1, \ldots, \psi_{k+1}\}$ han de ser ortogonals dos a dos, és a dir han de satisfer

$$0 = \langle \psi_j, \psi_l \rangle = \int_a^b w(x)\psi_j(x)\psi_l(x)dx \qquad \text{si} \qquad j \neq l.$$

Per tant, si es multiplica per $w(x)\psi_j(x)$ a banda i banda de (1.7) i s'integra entre a i b es té

$$\langle \phi_{k+1}, \psi_j \rangle = \left\langle \sum_{l=0}^{k+1} \alpha_l \psi_l, \psi_j \right\rangle = \sum_{l=0}^{k+1} \alpha_l \langle \psi_l, \psi_j \rangle = \alpha_j \langle \psi_j, \psi_j \rangle ,$$

la qual cosa determina els valors α_j per $j = 0 \div k$, concretament

$$\alpha_j = \frac{\langle \phi_{k+1}, \psi_j \rangle}{\langle \psi_j, \psi_j \rangle} = \frac{\langle x\psi_k, \psi_j \rangle}{\langle \psi_j, \psi_j \rangle} = \frac{\langle \psi_k, x\psi_j \rangle}{\langle \psi_j, \psi_j \rangle}$$

on a la darrera igualtat s'ha utilitzat que $\langle x\psi_k, \psi_j \rangle = \langle \psi_k, x\psi_j \rangle$, la qual cosa és directa si es reescriu la igualtat anterior utilitzant les integrals. Observem que $x\psi_j(x)$ és un polinomi de grau $j+1$, de

manera que si $j+1<k$ es té que $\alpha_j=0$ ja que, com es va veure, $\langle\psi_k,p_{k-1}\rangle=0$ si p_{k-1} és un polinomi de grau menor a k. Així, aïllant ψ_{k+1} a (1.7), es té

$$\begin{aligned}\psi_{k+1}(x)&=\frac{1}{\alpha_{k+1}}\left(\phi_{k+1}-\alpha_k\psi_k-\alpha_{k-1}\psi_{k-1}\right)=\frac{1}{\alpha_{k+1}}\left(x\psi_k(x)-\alpha_k\psi_k(x)-\alpha_{k-1}\psi_{k-1}(x)\right)\\&=\frac{1}{\alpha_{k+1}}\left((x-\alpha_k)\psi_k(x)-\alpha_{k-1}\psi_{k-1}(x)\right).\end{aligned}$$

Observem ara que el coeficient dominant de ψ_{k+1}, que denotem A_{k+1}, ha de ser igual al coeficient dominant del polinomi que es troba a la dreta de l'equació anterior, és a dir el coeficient que acompanya el terme x^{k+1}. Així, si A_k denota el coeficient dominant de ψ_k, es té que

$$A_{k+1}=\frac{1}{\alpha_{k+1}}A_k\qquad\Rightarrow\qquad\frac{1}{\alpha_{k+1}}=\frac{A_{k+1}}{A_k}.$$

Per tant, el polinomi ψ_{k+1} queda determinat si s'especifica A_{k+1} i es coneixen els polinomis ψ_k i ψ_{k-1}, ja que

$$\boxed{\psi_{k+1}=\frac{A_{k+1}}{A_k}\left((x-\alpha_k)\psi_k(x)-\alpha_{k-1}\psi_{k-1}(x)\right)}\tag{1.8}$$

amb

$$\alpha_k=\frac{\langle\psi_k,x\psi_k\rangle}{\langle\psi_k,\psi_k\rangle}\qquad\text{i}\qquad\alpha_{k-1}=\frac{\langle\psi_k,x\psi_{k-1}\rangle}{\langle\psi_{k-1},\psi_{k-1}\rangle}$$

Aquesta recurrència es coneix com a **recurrència dels polinomis ortogonals**. Observem que els dos primers polinomis ψ_0 (que és una constant de la forma A_0) i ψ_1 (que és un polinomi de la forma A_1x+a_0) s'han de calcular sense la recurrència. És fàcil veure, però, que estan determinats si A_0 i A_1 estan especificats, ja que aleshores $\psi_0(x)=A_0$ i

$$\psi_1(x)=A_1\left(x-\frac{\langle\psi_0,x\psi_0\rangle}{\langle\psi_0,\psi_0\rangle}\right)$$

on aquesta fórmula s'obtè imposant que $\langle\psi_1,\psi_0\rangle=0$. En particular observem que es pot obtenir ψ_1 aplicant la fórmula de recurrència dels polinomis ortogonals prenent $\psi_0=A_0$ i definint $\psi_{-1}=0$.

Observació. La fórmula de recurrència per a polinomis ortogonals permet concloure que una família de polinomis ortogonals respecte $[a,b]$ i una funció pes $w(x)$ està únivocament determinada si els coeficients dominants dels polinomis $\psi_0,\psi_1,\psi_2,\ldots$, que denotem $A_0,A_1,A_2,\ldots$ respectivament (i que no poden prendre el valor 0), estan prefixats.

Observació (Legendre). Analitzant quina forma prenen els termes α_k i α_{k-1} de la fórmula de recurrència dels polinomis ortogonals quan es consideren els polinomis de Legendre (estandaritzats), és a dir quan $[a,b]=[-1,1]$ i $w(x)=1$ i $\psi_k(1)=1$ per tot $k\geq0$, la fórmula es pot expressar de forma simplificada com

$$\boxed{\psi_{k+1}(x)=\frac{2k+1}{k+1}x\psi_k(x)-\frac{k}{k+1}\psi_{k-1}(x)\qquad\text{per}\quad k\geq1}\tag{1.9}$$

amb $\psi_0(x)=1$ i $\psi_1(x)=x$. Per a més informació consultar la bibliografia bàsica (per exemple Aubanell-Benseny-Delshams, pàgina 223). A problemes veureu que els polinomis de Legendre

també es poden obtenir amb la fórmula

$$\boxed{\psi_0(x) = 1 \qquad \text{i} \qquad \psi_k(x) = \frac{1}{2^k\, k!} \frac{d^k}{dx^k}[(x^2-1)^k] \qquad \text{per} \quad k \geq 1.} \tag{1.10}$$

Càlcul dels pesos de les fórmules Gaussianes

Quan es coneixen els polinomis ortogonals $\{\psi_0, \psi_1, \psi_2, \dots\}$ respecte un interval $[a,b]$ i una funció pes $w(x)$, així com els zeros $x_0, x_1, \dots, x_m$ del polinomi ψ_{m+1}, per aplicar la fórmula Gaussiana

$$\int_a^b w(x) f(x) dx \approx \sum_{k=0}^{m} W_k f(x_k)$$

només cal concretar el valor dels pesos W_k per $k = 0 \div m$. Sempre es podria calcular els pesos directament utilitzant la fórmula (1.5), però a continuació es dona una fórmula directa més pràctica.

Proposició. Els pesos W_k definits a (1.5) satisfan

$$\boxed{W_k = \frac{A_{m+1}}{A_m} \frac{\langle \psi_m, \psi_m \rangle}{\psi'_{m+1}(x_k)\psi_m(x_k)} \quad \text{per} \quad k = 0 \div m.} \tag{1.11}$$

Prova. Considerem la funció

$$g_k(x) = \frac{\psi_{m+1}(x)}{x - x_k} \psi_{m+2}(x).$$

que és un polinomi de grau $2m+2$ (noteu que $x - x_k$ és un factor de $\psi_{m+1}(x)$ ja qe x_k és un zero d'aquest polinomi) i satisfà $g_k(x_k) = \psi'_{m+1}(x_k)\psi_{m+2}(x_k)$ (utilitzeu la regla de l'Hôpital per resoldre la indeterminació) i $g_k(x_l) = 0$ per la resta de zeros de ψ_{m+1} (ja que $\psi_{m+1}(x_l) = 0$). De la fórmula d'integració Gaussiana (1.6) es té

$$\int_a^b w(x) g_k(x) dx - \sum_{l=0}^{m} W_l g_k(x_l) = \frac{g_k^{(2m+2)}(\xi)}{(2m+2)!} \frac{1}{A_{m+1}^2} \int_a^b w(x) \psi_{m+1}^2(x) dx.$$

Observem que:

$$\int_a^b w(x) g_k(x) dx = \int_a^b w(x) \frac{\psi_{m+1}(x)}{x - x_k} \psi_{m+2}(x) dx = \left\langle \frac{\psi_{m+1}(x)}{x - x_k}, \psi_{m+2}(x) \right\rangle = 0$$

perquè $\psi_{m+1}(x)/(x - x_k)$ és un polinomi de grau m (i s'ha vist que $\langle p, \psi_{m+2} \rangle = 0$ per tot polinomi p de grau igual o menor a $m+1$),

$$\sum_{l=0}^{m} W_l g_k(x_l) = W_k g_k(x_k)$$

perquè $g_k(x_l) = 0$ per tot $l \neq k$, i

$$\frac{g_k^{(2m+2)}(\xi)}{(2m+2)!} \frac{1}{A_{m+1}^2} \int_a^b w(x) \psi_{m+1}^2(x) dx = \frac{A_{m+2}}{A_{m+1}} \langle \psi_{m+1}, \psi_{m+1} \rangle ,$$

perquè la derivada $(2m+2)$-èssima d'un polinomi de grau $2m+2$ és igual al coeficient dominant del polinomi per $(2m+2)!$, i el coeficient dominant de $g_k(x)$ és $A_{m+1}A_{m+2}$. Així es pot aïllar W_k per obtenir

$$W_k = -\frac{A_{m+2}}{A_{m+1}}\frac{\langle \psi_{m+1}, \psi_{m+1}\rangle}{g_k(x_k)} = -\frac{A_{m+2}}{A_{m+1}}\frac{\langle \psi_{m+1}, \psi_{m+1}\rangle}{\psi'_{m+1}(x_k)\psi_{m+2}(x_k)},$$

i utilitzant la recurrència dels polinomis ortogonals, es pot reexpressar la fórmula anterior com

$$W_k = \frac{A_{m+1}}{A_m}\frac{\langle \psi_m, \psi_m\rangle}{\psi'_{m+1}(x_k)\psi_m(x_k)},$$

ja que de (1.8) es té clarament (recordeu que x_k és un zero de ψ_{m+1})

$$\psi_{m+2}(x_k) = -\frac{A_{m+2}}{A_{m+1}}\frac{\langle \psi_{m+1}, x\psi_m\rangle}{\langle \psi_m, \psi_m\rangle}\psi_m(x_k),$$

i aleshores

$$W_k = \frac{\langle \psi_{m+1}, \psi_{m+1}\rangle}{\langle \psi_{m+1}, x\psi_m\rangle}\frac{\langle \psi_m, \psi_m\rangle}{\psi'_{m+1}(x_k)\psi_m(x_k)},$$

i, d'altra banda, multiplicant a banda i banda de (1.8) per $w(x)\psi_{k+1}$ i integrant entre a i b es té

$$\langle \psi_{k+1}, \psi_{k+1}\rangle = \frac{A_{k+1}}{A_k}\langle \psi_{k+1}, x\psi_k\rangle,$$

de manera que prenent $k = m$ es té

$$\frac{\langle \psi_{m+1}, \psi_{m+1}\rangle}{\langle \psi_{m+1}, x\psi_m\rangle} = \frac{A_{m+1}}{A_m}$$

i la fórmula per W_k queda provada.

□

Observació (Legendre). Els pesos de la fórmula de Gauss-Legendre amb $m+1$ punts es poden expressar com

$$\boxed{W_k = \frac{2}{(1-x_k^2)\psi'_{m+1}(x_k)^2} \qquad \text{per} \qquad k = 0 \div m.}$$

Per més detalls consultar Aubanell-Benseny-Delshams, Problema IV.8 de la pàgina 392.

Exercici. a) Feu explícites la fórmula de Gauss-Legendre de 3 punts (i.e. $m+1=3$).

Com que $\psi_0(x) = 1$ i $\psi_1(x) = x$, a partir de la recurrència (1.9) obtenim

$$\psi_2(x) = \frac{3}{2}x^2 - \frac{1}{2}$$

i

$$\psi_3(x) = \frac{5}{2}x\left(x^2 - \frac{3}{5}\right).$$

Els zeros de ψ_3 són $x_0 = -\sqrt{15}/5$, $x_1 = 0$ i $x_2 = -x_0$, i aplicant els resultats generals anunciats per la fórmula de Gauss-Legendre es té

$$\int_{-1}^{1} f(x) = \frac{5}{9}f(x_0) + \frac{8}{9}f(x_1) + \frac{5}{9}f(x_2) + \frac{f^{(6)}(\xi)}{15750}$$

b) Utilitzeu la fórmula anterior per calcular la integral

$$\int_0^5 ye^{-0.1y^2}dy$$

i doneu un cota de l'error que es fa. Després compareu aquesta cota amb l'error real que es comet (per la qual cosa heu de resoldre la intergral exactament).

Fem un canvi de variable (lineal) per reescriure la integral sobre l'interval $[-1, 1]$. És a dir, busquem $x(y) = ay + b$ de manera que $x(0) = -1$ i $x(5) = 1$. Això determina $b = -1$ i $a = 2/5$. Amb aquest canvi de variables es té $y = (x - b)/a = 5(x + 1)/2$ i $dy = dx/a = 5/2dx$, i la intergral es reescriu com

$$\begin{aligned}\int_0^5 ye^{-0.1y^2}dy &= \int_{x(0)}^{x(5)} \frac{x-b}{a} e^{-0.1\left(\frac{x-b}{a}\right)^2} \frac{1}{a}dx = \int_{-1}^{1} \frac{5}{2}(x+1)e^{-\frac{75}{4}(x+1)^2}\frac{5}{2}dx \\ &= \frac{25}{4}\int_{-1}^{1}(x+1)e^{-0.1\frac{25}{4}(x+1)^2}dx = \frac{25}{4}\left(\frac{5}{9}f(x_0) + \frac{8}{9}f(x_1) + \frac{5}{9}f(x_2) + \frac{f^{(6)}(\xi)}{15750}\right)\end{aligned}$$

amb $f(x) = (x + 1)e^{-0.1\frac{25}{4}(x+1)^2}$. Fent el càlcul (ometent el terme d'error) ens surt l'aproximació

$$\int_0^5 ye^{-0.1y^2}dy \approx 4.59268...,$$

i el valor exacte és (la integral es pot calcular directament)

$$\int_0^5 ye^{-0.1y^2}dy = 4.58958...$$

de manera que es comet un error menor a 0.02. Si es fa una gràfica de $f^{(6)}$ entre -1 i 1, s'observa que $|f^{(6)}(x)| < 60$ per tot $x \in [-1, 1]$, la qual cosa ens permet acotar l'error de la fórmula gaussiana per

$$\frac{25}{4}\frac{|f^{(6)}(\xi)|}{15750} \leq \frac{25}{4}\frac{60}{15750} = 0.0238095,$$

la qual cosa està d'acord amb els resultats obtinguts.

c) Comproveu que l'aproximació de la intergral

$$\int_0^5 ye^{-\alpha y^2}dy$$

amb $\alpha > 0$ gran (per exemple $\alpha = 3$) és força dolenta. En aquest cas la derivada $f^{(6)}$ pren valors molt grans entre -1 i 1.

Fórmula de Gauss-Txebyshew

Considerem les funcions definides a $[-1,1]$ donades per

$$T_k(x) = \cos(k \arccos(x)).$$

Aquestes funcions són polinomis. Clarament

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= \cos(2\arccos(x)) = ? \end{aligned}$$

Per estudiar $T_2(x)$ definim $\alpha = \arccos(x)$ i utilitzem la fórmula de l'angle doble, de manera que

$$\cos(2\alpha) = \cos(\alpha)^2 - \sin(\alpha)^2 = \cos(\alpha)^2 - (1 - \cos(\alpha)^2) = 2\cos(\alpha)^2 - 1,$$

i desfent el canvi es té

$$T_2(x) = \cos(2\arccos(x)) = 2\cos(\arccos(x))^2 - 1 = 2x^2 - 1.$$

En general es té la recurrència

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \qquad \text{per} \qquad k \in \{1,2,3,...\}.$$

En efecte, utilitzant la fórmula pel cosinus de l'angle suma es té, per tot angle α,

$$\cos((k+1)\alpha)) = \cos(k\alpha)\cos(\alpha) - \sin(k\alpha)\sin(\alpha)$$

i

$$\begin{aligned} \cos((k-1)\alpha)) &= \cos(k\alpha)\cos(-\alpha) - \sin(k\alpha)\sin(-\alpha) \\ &= \cos(k\alpha)\cos(\alpha) + \sin(k\alpha)\sin(\alpha). \end{aligned}$$

Sumant aquestes dues equació s'obté

$$\cos((k+1)\alpha)) + \cos((k-1)\alpha)) = 2\cos(k\alpha)\cos(\alpha),$$

i prenent ara $\alpha = \arccos(x)$ es conclou

$$T_{k+1}(x) + T_{k-1}(x) = 2xT_k(x).$$

Per tant tots els polinomis de Txebyshew són polinomis.

Observem que de la recurrència anterior, i del fet que $T_0(x) = 1$ i $T_1(x) = x$, es dedueix que el coeficient dominant de $T_k(x)$ és 2^{k-1}.

Els polinomis de Txebyshew són una família de polinomis ortogonals respecte l'interval $[-1,1]$ i la funció pes $w(x) = 1/\sqrt{1-x^2}$. Per veure això cal comprovar que

$$\langle T_k, T_l \rangle = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} T_k(x) T_l(x) dx = 0 \quad \text{si} \quad k \neq l$$

i que $\langle T_k, T_k\rangle \neq 0$ per tot $k \in \{0,1,2,...\}$. Per veure-ho fem el canvi de variables $\alpha = \arccos(x)$ a la integral anterior:

$$\begin{aligned}\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} T_k(x)T_l(x)dx &= \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} \cos(k\arccos(x))\cos(l\arccos(x))dx \\ &= -\int_{\arccos(-1)}^{\arccos(1)} \cos(k\alpha)\cos(l\alpha)d\alpha = \int_0^{\pi} \cos(k\alpha)\cos(l\alpha)d\alpha\end{aligned}$$

ja que $\arccos'(x) = 1/\sqrt{1-x^2}$. D'aquí és clar que $\langle T_0, T_0\rangle = \pi$. D'altra banda, com que

$$\cos((k+l)\alpha)) = \cos(k\alpha)\cos(l\alpha) - \sin(k\alpha)\sin(l\alpha)$$

i

$$\begin{aligned}\cos((k-l)\alpha)) &= \cos(k\alpha)\cos(-l\alpha) - \sin(k\alpha)\sin(-l\alpha) \\ &= \cos(k\alpha)\cos(l\alpha) + \sin(k\alpha)\sin(l\alpha),\end{aligned}$$

sumant aquestes dues equacions es té que

$$\cos(k\alpha)\cos(l\alpha) = \frac{1}{2}\cos((k+l)\alpha)) + \frac{1}{2}\cos((k-l)\alpha)).$$

Observem finalment que, com $k+l$ és enter no nul si $k \in \{1,2,3,...\}$,

$$\int_0^{\pi} \cos((k+l)\alpha))d\alpha = \frac{1}{k+l}\left[\sin((k+l)\alpha)\right]_{\alpha=0}^{\alpha=\pi} = 0$$

i si $k-l \neq 0$, de manera que $k-l$ també és enter no nul,

$$\int_0^{\pi} \cos((k-l)\alpha))d\alpha = \frac{1}{k-l}\left[\sin((k-l)\alpha)\right]_{\alpha=0}^{\alpha=\pi} = 0$$

mentre que si $k = l$ es té

$$\int_0^{\pi} \cos(0)d\alpha = \int_0^{\pi} d\alpha = \pi.$$

Per tant es conclou $\langle T_k, T_l\rangle = 0$ si $k \neq l$ i

$$\langle T_k, T_k\rangle = \frac{\pi}{2} \qquad \text{si} \qquad k \in \{1,2,3,...\}$$

i $\langle T_0, T_0\rangle = \pi$.

Una de les avantatges dels polinomis de Txebyshew respecte els de Legendre és que els seus zeros són molt fàcils de calcular. En efecte, els zeros de $T_k(x)$ s'obtenen imposant

$$\begin{aligned}\cos(k\arccos(x)) = 0 \quad &\Rightarrow \quad k\arccos(x) = \frac{\pi}{2} + j\pi \qquad \text{per} \quad j \in \mathbb{Z} \\ &\Rightarrow \quad \arccos(x) = \frac{\pi(1+2j)}{2k} \qquad \text{per} \quad j \in \mathbb{Z} \\ &\Rightarrow \quad x = \cos\left(\frac{\pi(1+2j)}{2k}\right) \qquad \text{per} \quad j \in \mathbb{Z}\end{aligned}$$

Aquesta fórmula pren valors diferents si $j \in \{0, 1, 2, \ldots, k-1\}$, i cada un d'aquests valors és un zero simple de $T_k(x)$ (recordeu que per la teoria general de polinomis ortogonals sabem que n'hi ha tants com el grau del polinomi).

Per tant, lligant amb les fórmules Gaussianes, es té que els $m+1$ zeros de $T_{m+1}(x)$ són

$$x_k = \cos\left(\frac{\pi(1+2k)}{2(m+1)}\right) \qquad \text{per} \quad k \in \{0, 1, 2, 3, \ldots, m\}$$

i la fórmula de Gauss-Txebishew de $m+1$ punts pren la forma

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x) = \sum_{k=0}^{m} W_k f(x_k) + \text{Error} = \sum_{k=0}^{m} W_k f\left(\cos\left(\frac{\pi(1+2k)}{2(m+1)}\right)\right) + \text{Error}.$$

Els pesos W_k també són fàcils de calcular a partir de (1.11). Per aplicar aquesta fórmula, definim $\alpha_k = \arccos(x_k)$, i observem que

$$\begin{aligned} T'_{m+1}(x_k) &= (m+1)\cos'((m+1)\arccos(x_k))\arccos'(x_k) \\ &= (m+1)\sin((m+1)\arccos(x_k))\frac{1}{\sin(\arccos(x_k))} \\ &= (m+1)\sin((m+1)\alpha_k)\frac{1}{\sin(\alpha_k)}, \end{aligned}$$

on hem utilitzat que

$$\arccos'(x) = -\frac{1}{\sin(\arccos(x))}.$$

D'altra banda, aplicant una vegada més la fórmula de l'angle suma (amb els angles $(m+1)\alpha_k$ i $-\alpha_k$)

$$\begin{aligned} T_m(x_k) &= \cos(m\arccos(x_k)) \\ &= \cos(m\alpha_k) = \cos((m+1)\alpha_k)\cos(-\alpha_k) - \sin((m+1)\alpha_k)\sin(-\alpha_k) \\ &= \cos((m+1)\alpha_k)\cos(\alpha_k) + \sin((m+1)\alpha_k)\sin(\alpha_k) \\ &= \sin((m+1)\alpha_k)\sin(\alpha_k), \end{aligned}$$

on a la darrera igualtat hem utilitzat que $0 = \cos((m+1)\alpha_k) = \cos((m+1)\arccos(x_k)) = T_{m+1}(x_k)$ per ser x_k un zero de $T_{m+1}(x)$. Per tant, de la fórmula (1.11) i utilitzant que $A_{m+1}/A_m = 2$ i que $\langle T_m, T_m\rangle = \pi/2$, es té finalment

$$\begin{aligned} W_k &= \frac{\pi}{T'_{m+1}(x_k)T_m(x_k)} = \frac{\pi}{(m+1)\sin((m+1)\alpha_k)^2} = \frac{\pi}{(m+1)(1-\cos((m+1)\alpha_k)^2)} \\ &= \frac{\pi}{(m+1)(1-\cos((m+1)\arccos(x_k))^2)} = \frac{\pi}{(m+1)(1-T_{m+1}(x_k))} = \frac{\pi}{m+1} \end{aligned}$$

Observeu que els pesos W_k no depenen de l'índex k.

La senzillesa dels pesos W_k i dels zeros del polinomi $T_{m+1}(x)$ fa que sigui molt fàcil aplicar la fórmula de Gauss-Txebyshew amb un elevat nombre de punts. De fet es pot donar la fórmula explícita per un nombre de punts $m+1$ general:

$$\boxed{\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x) = \frac{\pi}{m+1}\sum_{k=0}^{m} f\left(\cos\left(\frac{\pi(1+2k)}{2(m+1)}\right)\right) + \text{Error},}$$

on a partir de (1.6) (i utilitzant que el coeficient dominant de T_{m+1} és $A_{m+1} = 2^m$ i que $\langle T_{m+1}, T_{m+1}\rangle = \pi/2$) es pot veure que l'error és

$$\boxed{\text{Error} = \frac{\pi}{2^{2m+1}(2m+2)!} f^{(2m+2)}(\xi) \qquad \text{amb} \quad \xi \in (-1,1).}$$

Observació. Per calcular la integral de $f(x)$ a l'interval $[-1,1]$, on f és contínua, podem reescriure la integral com

$$\int_{-1}^{1} f(x)dx = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} \sqrt{1-x^2} f(x)dx$$

i aplicar la fórmula de Gauss-Txebyshew a la funció $g(x) = \sqrt{1-x^2}f(x)$, que també és contínua a $[-1,1]$. Si f és $C^{2m+2}(-1,1)$, la funció g també ho serà i per tant es tindrà fórmula per l'error, però noteu que en els extrems de l'interval les derivades de g no estan acotades de manera que no es pot donar una cota de $g^{(2m+2)}$ sobre tot l'interval.

1.6 Mètodes d'integració adaptatius

Com s'ha comentat, establir l'error d'integració utilitzant cotes de les derivades de la funció que s'està integrant no és pràctic si es vol un algorisme automàtic. Per poder obtenir aproximacions d'una integral amb un error fixat a priori sense haver d'estudiar les derivades de la funció que s'integra cal recórrer als mètodes adaptatius. Aquests mètodes consisteixen en anar calculant aproximacions de la integral cada vegada més precises i en utilitzar les aproximacions que es van obtenint per estimar l'error que es comet.

Mètode de Romberg

Suposem que volem integrar la funció f entre a i b, i definim

$$J = \int_a^b f(x)dx.$$

Sigui $T_0(b-a)$ la fórmula dels trapezis, $T_0((b-a)/2)$ la fórmula dels trapezis composta amb dos subintervals, $T_0((b-a)/4)$ la fórmula dels trapezis composta amb 4 subintervals, i en general $T_0((b-a)/2^{n-1})$ la fórmula dels trapezis amb n subintervals.

Observem que si $x_{i+1} - x_i = h$ i $y_i = (x_i + x_{i+1})/2$ aleshores (utilitzant el desenvolupament de Taylor de f entorn de y_i)

$$\begin{aligned}\int_{x_i}^{x_{i+1}} f(x)dx &= \int_{x_i}^{x_{i+1}} f(y_i) + f'(y_i)(x-y_i) + \frac{1}{2}f''(y_i)(x-y_i)^2 + \frac{1}{6}f'''(y_i)(x-y_i)^3 + \frac{1}{24}f''''(y_i)(x-y_i)^4 + \dots\, dx \\ &= f(y_i)h + \alpha_2 h^3 f''(y_i) + \alpha_4 h^5 f''''(y_i) + \dots\end{aligned} \tag{1.12}$$

on α_2, α_4, etc. són coeficients que no ens cal conèixer. D'altra banda,

$$f(x_i) = f(y_i) - \frac{h}{2}f'(y_i) + \frac{h^2}{2^2}\frac{1}{2}f''(y_i) - \frac{h^3}{2^3}\frac{1}{6}f'''(y_i) + \frac{h^4}{2^4}\frac{1}{24}f''''(y_i) + ...$$
$$f(x_{i+1}) = f(y_i) + \frac{h}{2}f'(y_i) + \frac{h^2}{2^2}\frac{1}{2}f''(y_i) + \frac{h^3}{2^3}\frac{1}{6}f'''(y_i) + \frac{h^4}{2^4}\frac{1}{24}f''''(y_i) + ...$$

de manera que

$$\frac{f(x_i) + f(x_{i+1})}{2} = f(y_i) + \beta_2 h^2 f''(y_i) + \beta_2 h^4 f''''(y_i) + ...$$

i utilitzant aquesta relació a (1.12) es té

$$\int_{x_i}^{x_{i+1}} f(x)dx = h\frac{f(x_i) + f(x_{i+1})}{2} + (\alpha_2 - \beta_2)h^3 f''(y_i) + (\alpha_4 - \beta_4)h^5 f''''(y_i) + ...$$

Així, si $h = (b-a)/2^{n-1}$ i $x_i = a + ih$, es té que

$$\begin{aligned} J = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx &= T_0(h) + (\alpha_2 - \beta_2)h^3\sum_{i=0}^{n-1} f''(y_i) + (\alpha_4 - \beta_4)h^5\sum_{i=0}^{n-1} f''''(y_i) + ... \\ &= T_0(h) + (\alpha_2 - \beta_2)h^2\frac{b-a}{n}\sum_{i=0}^{n-1} f''(y_i) + (\alpha_4 - \beta_4)h^4\frac{b-a}{n}\sum_{i=0}^{n-1} f''''(y_i) + ... \\ &= T_0(h) + (\alpha_2 - \beta_2)h^2(b-a)f''(\xi) + (\alpha_4 - \beta_4)h^4(b-a)f''''(\eta) + ... , \end{aligned}$$

on s'ha utilitzat el teorema del valor mitjà per sumes (veure (1.3)) a la última igualtat. És a dir l'error de la fórmula de trapezis composta amb n subintervals es pot expressar com

$$J - T_0(h) = c_{0,0}h^2 + c_{0,1}h^4 + c_{0,2}h^6 + ... \tag{1.13}$$

on $c_{0,0}, c_{0,1}, c_{0,2}, ...$ són coeficients que no cal conèixer. Aquesta fórmula per l'error del mètode dels trapezis compost confirma que l'aproximació és d'ordre h^2 (cosa que ja havíem vist), però la sèrie completa en termes de potències de h serà útil per donar un mètode adaptatiu.

Observem que la relació (1.13) és vàlida per qualsevol h, per tant

$$\begin{aligned} J &= T_0(h) + c_{0,0}h^2 + c_{0,1}h^4 + c_{0,2}h^6 + ... \\ J &= T_0\left(\frac{h}{2}\right) + c_{0,0}\left(\frac{h}{2}\right)^2 + c_{0,1}\left(\frac{h}{2}\right)^4 + c_{0,2}\left(\frac{h}{2}\right)^6 + ..., \end{aligned}$$

de manera que si es multiplica la segona equació per 4 i es resta a la primera equació es té

$$J - 4J = T_0(h) - 4T_0\left(\frac{h}{2}\right) + (c_{0,1} - 4c_{0,1}/2^4)h^4 + (c_{0,2} - 4c_{0,2}/2^6)h^6 + ...,$$

i aïllant J es té

$$J = -\frac{1}{3}\left(T_0(h) - 4T_0\left(\frac{h}{2}\right)\right) + c_{1,1}h^4 + c_{1,2}h^6 + ...$$

Això equival a dir que la fórmula d'aproximació

$$T_1(h) = -\frac{1}{3}\left(T_0(h) - 4T_0\left(\frac{h}{2}\right)\right)$$

és d'ordre h^4 (millor que l'aproximació per trapezis), de fet es té que

$$J - T_1(h) = c_{1,1}h^4 + c_{1,2}h^6 + c_{1,3}h^8 + ...,$$

i podem repetir el procediment per definir

$$T_2(h) = \frac{1}{1-2^4}\left(T_1(h) - 2^4 T_1\left(\frac{h}{2}\right)\right)$$

que satisfa

$$J - T_2(h) = c_{2,2}h^6 + c_{2,3}h^8 + ...$$

En general es calcula $T_{j+1}(h)$ utilitzant la fórmula

$$T_{j+1}(h) = \frac{1}{1-4^j}\left(T_j(h) - 4^j T_j\left(\frac{h}{2}\right)\right)$$

i es té que

$$J - T_{j+1}(h) = c_{j+1,j+1}h^{2(j+2)} + ...$$

Obserevem finalment que, si h és petit,

$$\begin{aligned} T_{j+1}(h) - T_j(h) &= -(J - T_{j+1}(h)) + (J - T_j(h)) = c_{j,j}h^{2(j+1)} + (-c_{j+1,j+1} + c_{j,j+1})h^{2(j+2)} + ... \\ &\approx c_{j,j}h^{2(j+1)} \approx J - T_j(h), \end{aligned}$$

i per tant dos iterats consecutius $T_{j+1}(h)$ i $T_j(h)$ ens permeten estimar l'error de l'aproximació j-èssima.

El mètode de Romberg consisteix, doncs, en calcular la successió $T_0(h)$, $T_1(h)$, $T_2(h)$, etc. fins que l'error entre dos iterats consecutius sigui menor a una tolerància $\varepsilon > 0$ fixada, és a dir fins que $|T_{j+1}(h) - T_j(h)| < \varepsilon$. Els passos del mètode de Romberg són els següents:

1. càlcul de $T_0(b-a)$,
2. càlcul de $T_0((b-a)/2)$ i $T_1(b-a)$,
3. càclul de $T_0((b-a)/4)$ i $T_1((b-a)/2)$ i $T_2(b-a)$,
4. càclul de $T_0((b-a)/8)$ i $T_1((b-a)/4)$ i $T_2((b-a)/2)$ i $T_3(b-a)$,
5. etc.

fins que se satisfaci $|T_{j+1}(h) - T_j(h)| < \varepsilon$. Observeu que per calcular $T_0((b-a)/2^{j+1})$ es pot aprofitar el càlcul de $T_0((b-a)/2^j)$, ja que la meitat dels termes del sumatori de $T_0((b-a)/2^{j+1})$ estan presents al sumatori de $T_0((b-a)/2^j)$ (i és absurd tornar-los a calcular). Concretament es té la recurrència

$$T_0\left(\frac{b-a}{2^{j+1}}\right) = \frac{T_0\left(\frac{b-a}{2^j}\right)}{2} + \frac{b-a}{2^{j+1}}\sum_{i=0}^{2^j-1} f(\bar{x}_i) \qquad \text{on} \qquad \bar{x}_i = \frac{x_i + x_{i+1}}{2} \qquad \text{i} \quad x_i = a + i\frac{b-a}{2^j}.$$

.

Mètode estàndard (per fórmules de Newton-Cotes)

Considerem la fórmula de Newton-Cotes amb $m+1$ nodes (és a dir m intervals), que és d'ordre

$$p = \begin{cases} m & \text{si } m \text{ senar} \\ m+1 & \text{si } m \text{ parell} \end{cases}$$

Denotem per $\mathrm{NC}_m(f,[a,b])$ el resultat d'aplicar la fórmula de Newton-Cotes a la funció amb $m+1$ punts a la funció f sobre l'interval $[a,b]$, i recordem que l'error ve donat per

$$\int_a^b f(x)dx - \mathrm{NC}_m(f,[a,b]) = C_m f^{(p+1)}(\xi)h^{p+2}$$

amb $h=(b-a)/m$, $\xi \in (a,b)$ i $C_m = K_m/(p+1)!$ amb K_m una constant independent de f i de l'interval $[a,b]$.

Suposem que volem calcular la integral

$$I = \int_a^b f(x)dx$$

amb una precisió donada ε. A continuació veurem que, definint

$$Q_n := \sum_{i=0}^{2^n-1} \mathrm{NC}_m(f,[x_i,x_{i+1}]) \qquad \text{amb} \qquad x_i = a + (b-a)\frac{i}{2^n}$$

per $n \in \{0,1,2,\dots\}$, que no és altra cosa que la fórmula composta de Newton-Cotes de $m+1$ sobre 2^n subintervals de $[a,b]$, aleshores es té que

$$I - Q_{n+1} = \frac{1}{2^{p+1}-1}(Q_{n+1}-Q_n) + o\left(\frac{1}{2^{(p+1)n}}\right), \tag{1.14}$$

on el terme $o(1/2^{(p+1)n})$ tendeix a zero quan $n \to \infty$, la qual cosa ens diu que podem utilitzar l'aproximació

$$I - Q_{n+1} \approx \frac{1}{2^{p+1}-1}(Q_{n+1}-Q_n)$$

Així, si el terme de la dreta de l'aproximació anterior és menor que ε, es podrà dir que l'error entre Q_{n+1} i la integral és menor a ε. Per tant, la idea és anar calculant $Q_0, Q_1, Q_2, \dots$ fins que el valor de dos iterats consecutius satisfaci

$$\frac{1}{2^{p+1}-1}|Q_{n+1}-Q_n| < \varepsilon,$$

ja que aleshores Q_{n+1} serà l'aproximació de la integral amb la precisió demanada.

Vegem que se satisfà (1.14). Per simplificar la notació denotem

$$Q_n^{(i)} := \mathrm{NC}_m(f,[x_i,x_{i+1}]),$$

que es corresponen amb els sumands de l'expressió que defineix Q_n. Observem que els sumands de Q_{n+1} s'obtenen quan es divideix per 2 cada interval $[x_i, x_{i+1}]$ i s'aplica la fórmula de Newton-Cotes en cada subinterval. Per tant Q_{n+1} es pot escriure com

$$Q_{n+1} = \sum_{i=0}^{2^n-1} Q_{n,\text{ref}}^{(i)}$$

amb

$$Q_{n,\text{ref}}^{(i)} := \text{NC}_m(f, [x_i, \bar{x}_i]) + \text{NC}_m(f, [\bar{x}_i, x_{i+1}]) \quad \text{amb} \quad x_i = a + (b-a)\frac{i}{2^n} \quad \text{i} \quad \bar{x}_i = \frac{x_i + x_{i+1}}{2},$$

que equival a l'aproximació de la integral sobre l'interval $[x_i, x_{i+1}]$ després de refinar aquest interval en dos subintervals. Definim

$$I^{(i)} = \int_{x_i}^{x_{i+1}} f(x)dx.$$

Noteu que és suficient veure la relació

$$I^{(i)} - Q_{n,\text{ref}}^{(i)} = \frac{1}{2^{p+1}-1}(Q_{n,\text{ref}}^{(i)} - Q_n^{(i)}) + o\left(\frac{1}{2^{(p+2)n}}\right) \tag{1.15}$$

a l'interval arbitrari $[x_i, x_{i+1}]$, ja que aleshores

$$\begin{aligned} I - Q_{n+1} &= \sum_{i=0}^{2^n-1} I^{(i)} - Q_{n,\text{ref}}^{(i)} = \sum_{i=0}^{2^n-1} \left(\frac{1}{2^{p+1}-1}(Q_{n,\text{ref}}^{(i)} - Q_n^{(i)}) + o\left(\frac{1}{2^{(p+2)n}}\right)\right) \\ &= \frac{1}{2^{p+1}-1}(Q_{n+1} - Q_n) + 2^n o\left(\frac{1}{2^{(p+2)n}}\right) \\ &= \frac{1}{2^{p+1}-1}(Q_{n+1} - Q_n) + o\left(\frac{1}{2^{(p+1)n}}\right). \end{aligned}$$

Per veure (1.15), definim

$$h_n = \frac{h}{2^n} = \frac{b-a}{m2^n},$$

que es correspon amb la longitud de l'interval $[x_i, x_{i+1}]$ dividida per m, i observem d'una banda que (desenvolupant per Taylor $f^{(p+1)}(\xi)$ entorn de $\bar{x}_i$)

$$I^{(i)} - Q_n^{(i)} = C_m f^{(p+1)}(\xi) h_n^{p+2} = C_m f^{(p+1)}(\bar{x}_i) h_n^{p+2} + o(h_n^{p+2}),$$

ja que $|\xi - \bar{x}_i| < m h_n$ (perquè $\xi \in (x_i, x_{i+1})$) i aleshores $h_n^{p+2}(\xi - \bar{x}_i)$ és $o(h_n^{p+2})$.

D'altra banda (aplicant novament el desenvolupament de Taylor dues vegades)

$$\begin{aligned} I^{(i)} - Q_{n+1}^{(i)} &= \int_{x_i}^{\bar{x}_i} f(x)dx - \text{NC}_m(f, [x_i, \bar{x}_i]) + \int_{\bar{x}_i}^{x_{i+1}} f(x)dx - \text{NC}_m(f, [\bar{x}_i, x_{i+1}]) \\ &= C_m f^{(p+1)}(\xi_1) \left(\frac{h_n}{2}\right)^{p+2} + C_m f^{(p+1)}(\xi_2) \left(\frac{h_n}{2}\right)^{p+2} \\ &= C_m f^{(p+1)}(\bar{x}_i) \left(\frac{h_n}{2}\right)^{p+2} + o(h_n^{p+2}) + C_m f^{(p+1)}(\bar{x}_i) \left(\frac{h_n}{2}\right)^{p+2} + o(h_n^{p+2}) \\ &= \frac{1}{2^{p+1}} C_m f^{(p+1)}(\bar{x}_i) h_n^{p+2} + o(h_n^{p+2}). \end{aligned}$$

Combinant aquestes dues expressions s'obté la relació

$$I^{(i)} - Q_{n+1}^{(i)} = \frac{1}{2^{p+1}}(I^{(i)} - Q_n^{(i)}) + o(h_n^{p+2}).$$

Aïllant $I^{(i)}$ de l'equació anterior es té

$$I^{(i)} = \frac{2^{p+1}Q_{n+1}^{(i)} - Q_n^{(i)}}{2^{p+1}-1} + o(h_n^{p+2})$$

de manera que

$$\begin{aligned} I^{(i)} - Q_{n+1}^{(i)} &= \frac{2^{p+1}Q_{n+1}^{(i)} - Q_n^{(i)}}{2^{p+1}-1} + o(h_n^{p+2}) - Q_{n+1}^{(i)} \\ &= \frac{1}{2^{p+1}-1}(Q_{n+1}^{(i)} - Q_n^{(i)}) + o(h_n^{p+2}), \end{aligned}$$

i com que $o(h_n^{p+2}) = o(1/2^{(p+2)n})$, la relació (1.15) queda provada.

Observació. Com s'ha comentat al final de la secció on s'explica el mètode de Romberg, quan s'aplica el procediment anterior utilitzant la fórmula dels trapezis, per calcular Q_{n+1} podem aprofitar el terme Q_n ja que la meitat dels termes del sumatori de Q_{n+1} estan presents al sumatori de Q_n (i és absurd tornar-los a calcular ja que alenteix el temps de comput). Concretament es té la recurrència

$$Q_{n+1} = \frac{Q_n}{2} + \frac{b-a}{2^{n+1}} \sum_{i=0}^{2^n-1} f(\bar{x}_i) \qquad \text{on} \qquad \bar{x}_i = \frac{x_i + x_{i+1}}{2} \qquad \text{i} \quad x_i = a + i\frac{b-a}{2^n}$$

amb

$$Q_0 = \frac{b-a}{2}(f(a) + f(b)).$$

Problema. El procediment anterior va refinant els diferents intervals de forma homogènia fins a arribar a la tolerància demanada. Un procediment més eficient assignaria una tolerància a cada interval proporcional a la seva longitud, i el refinament es faria només en els intervals en els que la tolerància encara no s'hagués assolit. Per exemple, donada una tolerància ε, es calcula Q_0 i Q_1 com en el cas anterior i suposem

$$\frac{1}{2^{p+1}-1}|Q_1 - Q_0| > \varepsilon,$$

de manera que el refinament ha de continuar. Ara tenim dos intervals de longitud $(b-a)/2$, concretament els intervals

$$\left[a, \frac{a+b}{2}\right] \qquad \text{i} \qquad \left[\frac{a+b}{2}, b\right].$$

La idea és assignar ara la tolerància $\varepsilon/2$ a cada un d'aquest intervals, i refinar-los fins assolir aquesta tolerància. Segons la notació anterior, l'aproximació al primer subinterval és $Q_1^{(0)}$ i la del segon $Q_1^{(1)}$. Així començaríem pel primer interval i calcularíem la integral refinada en aquest interval, que seria $Q_{1,\text{ref}}^{(0)}$. Si

$$\frac{1}{2^{p+1}-1}|Q_{1,\text{ref}}^{(0)} - Q_1^{(0)}| \leq \frac{\varepsilon}{2},$$

aleshores a l'interval $[a,(a+b)/2]$ prendríem l'aproximació $Q^{(0)}_{1,\text{ref}}$ i ja no caldría refinar-lo més (en cas contrari hauríem de seguir refinant, però per fixar idees suposem que no cal). Passaríem aleshores a considerar el segon interval, i calcularíem novament la integral refinada en aquest interval, que seria $Q^{(1)}_{1,\text{ref}}$. Si

$$\frac{1}{2^{p+1}-1}|Q^{(1)}_{1,\text{ref}} - Q^{(1)}_{1}| > \frac{\varepsilon}{2},$$

aleshores $Q^{(1)}_{1,\text{ref}}$ no seria una aproximació amb la precisió demanada a l'interval $[(a+b)/2,b]$ i hauríem de refinar els dos subintervals en els que s'ha dividit aquest interval, és a dir els intervals

$$\left[\frac{a+b}{2},(a+b)\frac{3}{4}\right] \qquad \text{i} \qquad \left[(a+b)\frac{3}{4},b\right].$$

A cada subinterval assignaríem la tolerància $\varepsilon/4$ i els refinaríem fins assolir aquesta tolerància seguint el mateix procediment.

Per pensar. Escriviu un codi que, donada la tolerància ε, calculi la integral amb aquesta tolerància realitzant refinaments no homogenis com s'acaba d'il·lustrar (utilitzeu la fórmula dels trapezis). Si guardeu la informació sobre els refinaments que es fan, al final del procés podeu graficar en quines zones s'han utilitzat més punts i en quines menys. Podeu posar-lo a prova amb la funció $f(x) = xe^{-7x}$ i l'interval $[a,b] = [0,2]$.

1.7 Integrals singulars

El fórmules d'integració interpolatòria són vàlides per funcions contínues en intervals tancats $[a,b]$. Quan la funció que es vol integrar presenta singularitats, cal fer un tractament previ de la integral per tal d'aplicar les fórmules interpolatòries satisfactòriament. Vegem algunes de les situacions que ens podem trobar i com gestionar-les.

- Funció f que té una discontinuïtat de salt en $c \in (a,b)$. En aquests casos, si f és regular als subintervals $[a,c]$ i $[c,b]$, aleshores cal treballar per separat amb les integrals

$$I_1 = \int_a^c f(x)dx \qquad \text{i} \qquad I_2 = \int_c^b f(x)dx.$$

- Funció f que tendeix a infinit en a, concretament funcions que són de la forma

$$f(x) = \frac{\phi(x)}{(x-a)^\mu}$$

amb $\mu \in [0,1)$ i $\phi \in C^1([a,b])$. Noteu que aquestes propietats garanteixen que la integral de $f(x)$ en $[a,b]$ és finita. En aquest cas es pot reescriure la integral com

$$\int_a^b f(x)dx = \int_a^b \frac{\phi(x)}{(x-a)^\mu}dx = \int_a^b \frac{\phi(x)-\phi(a)}{(x-a)^\mu}dx + \int_a^b \frac{\phi(a)}{(x-a)^\mu}dx.$$

Aleshores la segona integral es pot resoldre de forma exacte (ja que coneixem la primitiva de $(x-a)^{-\mu}$) i la primera integral es pot resoldre numèricament sense problemes ja que l'integrand és una funció contínua en $[a,b]$. En efecte, pel teorema del valor mitjà es té que $\phi(x)-\phi(a)=\phi'(\xi(x))(x-a)$, de manera que

$$\frac{\phi(x)-\phi(z)}{(x-a)^{\mu}}=\phi'(\xi(x))(x-a)^{1-\mu}$$

que clarament no presenta la singularitat a a si $\mu\in[0,1)$.

- Integral en un interval no acotat. Si es vol calcular

$$\int_a^{\infty} f(x)dx,$$

de la qual se sap que té un valor finit, cal trobar $c>a$ de manera que

$$\int_c^{\infty} f(x)dx<\frac{\delta}{2}$$

i després calcular numèricament la integral

$$\int_a^{c} f(x)dx$$

amb un error menor a $\delta/2$. D'aquesta manera l'error entre l'aproximació de l'integral a l'interval $[a,c]$ i la integral real sobre $[a,\infty)$ serà menor a δ.

Interpretació de la ortoganilitat

L'espai de funcions continues reals definides sobre un interval $[a,b]$ té estructura d'espai vectorial sobre $\mathbb{R}$ (la suma de funcions continues és una funció continua i el producte d'una funció contínua per un real és una funció contínua). A diferència de $\mathbb{R}^n$, la dimensió d'aquest espai és infinita, però és possible donar subespais d'aquest espai que tinguin dimensió finita. Per exemple, els polinomis de grau menor o igual a m és un subespai d'aquest espai funcional que té dimensió $m+1$ (ja que qualsevol polinomi de grau m o menor es pot expressar com a combinació lineal dels polinomis 1, x, x^2, $\ldots$, x^m).

Donada $w:[a,b]\to\mathbb{R}_+$ una funció pes positiva, es pot definir sobre aquest espai de funcions el producte escalar:

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

Aleshores, si $\{\psi_0,\psi_1,\psi_2,\ldots\}$ és una família de polinomis ortogonals respecte $[a,b]$ i w, es té que la projecció d'una funció f sobre el subespai generat per $\{\psi_0,\psi_1,\psi_2,\ldots,\psi_m\}$ és el polinomi de grau m o menor que minimitza

$$\langle f-p,f-p\rangle\,.$$

La projecció de f sobre el subespai generat per $\{\psi_0, \psi_1, \psi_2, \ldots, \psi_m\}$, que denotem per $p_m(f)$, és anàloga a com es fa a $\mathbb{R}^n$, concretament

$$p_m(f) = \sum_{k=0}^{m} \frac{\langle f, \psi_k \rangle}{\langle \psi_k, \psi_k \rangle} \psi_k.$$

Es pot veure que $p_m(f)$ així definit coincideix amb el la interpolació de f sobre els zeros del polinomi ψ_{m+1}, és a dir que

$$p_m(f) = \sum_{k=0}^{m} f(x_k) l_k$$

on x_k és el zero k-èssim de ψ_{m+1} i l_k és el polinomi que val 0 sobre x_j per $j \neq k$ i val 1 sobre x_k.

Capítol 2

Mètodes de Montecarlo

Bibliografia. Per aprofundir i ampliar els continguts d'aquest tema es pot consultar qualsevol llibre que tracti sobre els mètodes de Montecarlo. Per elaborar aquestes notes s'ha seguit el material del professor Tom Kennedy publicat a `https://www.math.arizona.edu/~tgk/mc/` que al seu torn segueix el llibre del professor Art Owen disponible a `https://artowen.su.domains/mc`.

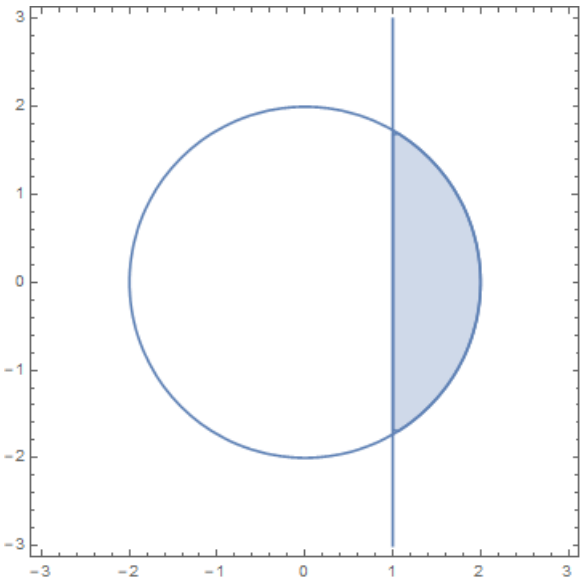
2.1 Base teòrica dels mètodes de Montecarlo

Els mètodes anomenats de Montecarlo permeten calcular determinades quantitats de forma probabilística. Vegem primer un parell d'exemples per il·lustrar el mètode per després desenvolupar la teoria d'una forma més general.

Exemple 1. Considerem el conjunt del pla donat per

$$C = \{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 < 4 \quad \text{i} \quad x > 1\}$$

que gràficament es correspon a l'àrea ombrejada del següent gràfic:



Observem que aquest conjunt està inclòs en el rectangle

$$R = \{(x, y) \in \mathbb{R}^2 \,|\, 1 < x < 2 \quad \text{i} \quad -2 < y < 2\},$$

del qual coneixem l'àrea, $\text{Àrea(R)} = 1 \cdot 4 = 4$.

Una manera probabilística de calcular l'àrea de C consisteix en generar n punts aleatoris amb distribució uniforme sobre el rectangle R i comptar quants d'aquest punts es troben dins del conjunt C. Com que la probabilitat que cada punt es trobi dins de C és igual a la fracció entre l'àrea de C i l'àrea de R, es té que

$$\frac{\text{Àrea}(C)}{\text{Àrea}(R)} \approx \frac{c}{n} \qquad \Rightarrow \qquad \text{Àrea}(C) \approx 4\frac{c}{n}$$

on c denota el nombre de punts que es troben dins del conjunt C (dels n punts generats). Així, podríem trobar l'àrea de C implementant el pseudocodi

```
c=0;
n=1000;
for (i = 1, i < n+1, i++) {
    x = Unif(1,2);
    y = Unif(-2,2);
    if (x^2 + y^2 < 4) {
        c++;
    }
}
areaC = 4 * c / n;
```

on la funció `Unif(a,b)` retorna un nombre aleatori amb distribució uniforme sobre l'interval (a, b).

Observem que el procediment que hi ha rere l'algorisme consisteix en calcular la mitjana empírica d'una variable aleatòria, concretament la variable

$$Z = 4\mathbb{1}_C(X, Y)$$

on $X \sim U(1, 2)$ i $Y \sim U(-2, 2)$ són uniformes sobre els intervals indicats, i on $\mathbb{1}_C(x, y) = 1$ si $(x, y) \in C$ i $\mathbb{1}_C(x, y) = 0$ altrament (es diu que la funció $\mathbb{1}_C$ és la funció indicador sobre el conjunt C). Per tant, l'aproximació que dona l'algorisme per l'àrea de C tindrà sentit sempre que l'esperança de Z sigui exactament l'àrea de C (perquè la mitjana empírica amb n realitzacions de la variable Z és un estimador de $\mathbb{E}(Z)$). Anem a comprovar que això és així (fins ara només ho hem "intuït" o "anticipat", però no ho hem provat formalment), és a dir vegem que

$$\mathbb{E}(Z) = \text{Àrea}(C).$$

Com que Z només pren dos valors (0 o 4), la seva esperança és

$$\mathbb{E}(Z) = 0 \cdot P(Z = 0) + 4 \cdot P(Z = 4) = 4 \cdot P(Z = 4),$$

per tant cal calcular simplement $P(Z = 4)$, que equival a la probabilitat que el vector (X, Y) sigui un element de C, és a dir que $P((X, Y) \in C)$. Aquesta probabilitat es pot expressar com

$$\int_C f_{(X,Y)}(x, y)dxdy,$$

on $f_{(X,Y)}$ denota la funció de densitat del vector aleatori (X, Y). Cal, per tant, conèixer la densitat de (X, Y). Com que les components del vector (X, Y) són independents, es té que

$$f_{(X,Y)}(x, y) = f_X(x)f_Y(y)$$

on f_X és la densitat de X i f_Y és la densitat de Y. Com que $X \sim U(1, 2)$ i $Y \sim U(-2, 2)$, es té

$$f_{(X,Y)}(x, y) = f_X(x)f_Y(y) = \mathbb{1}_{[-1,2]}(x)\frac{1}{4}\mathbb{1}_{[-2,2]}(y).$$

Així es conclou finalment que

$$\begin{aligned}\mathbb{E}(Z) = 4 \cdot P(Z = 4) = 4 \cdot P((X, Y) \in C) = 4\int_C f_{(X,Y)}(x, y)dxdy \\ = 4\int_C \mathbb{1}_{[-1,2]}(x)\frac{1}{4}\mathbb{1}_{[-2,2]}(y)dxdy = \int_C dxdy = \text{Àrea}(C),\end{aligned}$$

on a la penúltima igualtat s'ha utilitzat que $C \subset [1, 2] \times [-2, 2]$ (de manera que l'integrand val 1 en tot el domini d'integració), i a l'última igualtat s'ha utilitzat que la integral de la funció constant igual a 1 sobre un domini és igual a la mesura del domini (i la mesura és l'àrea si el domini d'integració és un conjunt de $\mathbb{R}^2$).

Exemple 2. Considerem la integral

$$I = \int_a^b g(x)dx,$$

Una manera de calcular aquesta integral consisteix en generar n punts amb distribució uniforme sobre l'interval (a, b), que denotem per $x_1, x_2, \ldots, x_n$, i fer la mitjana dels valors $g(x_1), g(x_2), \ldots, g(x_n)$,

ja que aquesta mitjana multiplicada per $b-a$ és una aproximació de I si n és prou gran i el valor de I és finit. És a dir, es té l'aproximació

$$\int_a^b g(x)dx \approx (b-a)\frac{1}{n}\sum_{i=1}^{n} g(x_i). \tag{2.1}$$

Aquesta aproximació es basa en que, si X és una variable aleatòria amb densitat f_X, aleshores l'esperança de $g(X)$, que és una altra variable aleatòria, satisfà

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

Així, si X és una variable aleatòria uniformement distribuïda a (a,b), la seva funció de densitat és

$$f_X(x) = \frac{1}{b-a}\mathbb{1}_{(a,b)}(x)$$

on $\mathbb{1}_B$ és la **funció indicador** sobre B, és a dir

$$\mathbb{1}_B(x) = \begin{cases} 1 & \text{si } x \in B \\ 0 & \text{si } x \notin B \end{cases},$$

de manera que

$$\mathbb{E}(g(X)) = \frac{1}{b-a}\int_a^b g(x)dx.$$

Per tant, com que l'esperança $\mathbb{E}(g(X))$ es pot aproximar com

$$\mathbb{E}(g(X)) \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i),$$

l'aproximació (2.1) queda justificada.

Amb aquests exemples hem vist la **base del mètode de Montecarlo**:

Si una quantitat μ es pot expressar com l'esperança d'una variable aleatòria X, aleshores la quantitat μ en qüestió es pot aproximar si sabem generar realitzacions de la variable aleatòria X.

La metodologia per fer-ho es fonamenta en el següent teorema:

Llei dels grans nombres. *Siguin $X_1, X_2, X_3, \ldots$ variables aleatòries idènticament distribuïdes tals que $\mathbb{E}(X_i) = \mu$ per tot $i \in \mathbb{N}$. Aleshores per tot $\varepsilon > 0$ es té que*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) = 0.$$

Aquest teorema ens diu, per tant, que donat $\varepsilon > 0$, podem prendre n prou gran tal que, per exemple,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) < 0.01,$$

que equival a dir que

$$P\left(\mu \in (\mu_n - \varepsilon, \mu_n + \varepsilon)\right) \geq 1 - 0.01 = 0.99,$$

on μ_n és la variable aleatòria

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.2}$$

En altres paraules: si es pren una realització $\hat{\mu}_n$ de μ_n a l'atzar, és a dir es pren

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

amb x_i una realització de X_i, es té que el valor buscat μ està a l'interval

$$(\hat{\mu}_n - \varepsilon, \hat{\mu}_n + \varepsilon) \tag{2.3}$$

amb un grau de confiança del 99% (es pot dir amb una certesa del 99% que μ es troba en aquest interval). En general si es vol que

$$\mu \in (\hat{\mu}_n - \varepsilon, \hat{\mu}_n + \varepsilon)$$

amb un grau de confiança de $1 - \alpha$, aleshores cal prendre un n prou gran de manera que

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \varepsilon\right) < \alpha.$$

Cal remarcar un punt important aquí: un mètode probabilístic, com el de Montecarlo, no ens pot donar mai el resultat que busquem amb una certesa absoluta. Si α és molt petit podem estar molts segurs que el valor μ estarà dins de l'interval de confiança (2.3), però no absolutament segurs. Això és radicalment diferent als mètodes deterministes, ja que aquests mètodes, si s'apliquessin per calcular μ, si permetrien dir que μ està en un determinat interval amb una seguretat del 100% (per exemple, els mètodes que vam veure per calcular integrals en el tema 1 ens permetrien dir que el valor de la intergral es troba segur a a l'interval amb centre el valor calculat pel mètode i amb radi igual a una cota superior de l'error que comet el mètode).

El teorema anterior, però, no ens permet dir com de gran ha de ser n en funció d'α i d'ε.

Per abordar aquesta qüestió cal saber com es comporten les variables aleatòries μ_n a mesura creix n. Com que μ_n és suma de variables aleatòries independents, la variància de μ_n és igual a la suma de les variàncies de les variables que s'estan sumant. Com que, a més a més, les X_i per $i = 1 \div n$ són idènticament distribuïdes, si les seves variàncies són iguals a σ^2, és a dir si $\mathrm{Var}(X_i) = \sigma^2$ per $i = 1 \div n$, aleshores es conclou que

$$\mathrm{Var}(\mu_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

La variància dóna l'esperança del quadrat de la distància entre una variable aleatòria i la seva mitjana. Així, pel cas concret en que la variable és μ_n, com que la seva mitjana és μ, es té

$$\mathrm{Var}(\mu_n) = \mathbb{E}((\mu_n - \mu)^2),$$

i per tant la distància entre μ_n i μ és d'ordre $\sigma/\sqrt{n}$.

Tornem a veure, doncs, que com més gran sigui n més a prop estarà μ de les realitzacions de μ_n, però amb aquest anàlisi una mica més fi hem pogut determinar com millora l'aproximació a mesura augmentem n. Concretament si volem que l'esperança de la distància entre μ_n i el valor μ es redueixi a la meitat, cal multiplicar per 4 el valor de n, és a dir cal considerar l'estimador μ_{4n}. Per al càlcul d'integrals definides sobre un interval de la recta real aquesta velocitat de convergència és relativament baixa en comparació als mètodes deterministes estudiats al tema 1 (com a comparativa observeu que multiplicar per 4 el nombre d'intervals quan s'utilitza la fórmula dels trapezis composta implica que l'error de l'aproximació es divideix per 16). Amb integrals definides sobre subconjunts del pla o de l'espai, però, multiplicar per 4 el nombre d'intervals en cada direcció espacial implica multiplicar per 4^2 o per 4^3 el nombre d'operacions que es fan per aproximar la integral, i per tant en aquests casos s'hauria de comparar com millora el mètode de Montecarlo per calcular aquestes integrals quan el nombre de punts per calcular la mitjana es multiplica per 4^2 o 4^3 respectivament. Una de les virtuts del mètode de Montecarlo és que la velocitat de convergència és d'ordre $\sigma/\sqrt{n}$ independentment de si la integral és 1-dimensional o multidimensional, i per tant multiplicar per 4^k el nombre de punts que es fan servir per calcular la mitjana que aproxima la quantitat μ que es busca implica multiplicar per 2^k la precisió de l'aproximació. Veiem com el mètode de Montecarlo comença a ser competitiu per calcular integrals definides en dimensió no massa petita (de fet, si donem per fet que la convergència d'una fòrmula dels trapezis generalitzada a dimensions més alta es comporta com en el cas 1-dimensional, en el sentit que multiplicar per 4 el nombre d'intervals en cada dimensió comporta multiplicar per 16 la precisió de l'aproximació, aleshores la velocitat de convergència del mètode de Montecarlo en dimensió 4 seria anàloga a la que s'obtindria utilitzant la fòrmula dels trapezis, ja que $2^4 = 16$, i si la dimensió és superior a 4 el mètode de Montecarlo convergiria més ràpidament que el mètode determinista).

Tot i que l'observació anterior ens indica com es comporta l'aproximació a mesura s'augmenta n no ens permet encara dir quin n hem de prendre si volem un determinat grau de confiança en relació a la quantitat que estem aproximant. Per donar resposta a aquesta qüestió ens cal un segon teorema:

Teorema central del límit. *Siguin $X_1, X_2, X_3, \ldots$ variables aleatòries idènticament distribuïdes tals que $\mathbb{E}(X_i) = \mu$ i $Var(X_i) = \sigma^2 < \infty$ per tot $i \in \mathbb{N}$. Definim la variable aleatòria*

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Aleshores

$$Z_n = \frac{\sqrt{n}}{\sigma}(\mu_n - \mu)$$

convergeix en llei a una variable aleatòria normal amb mitjana 0 i variància 1. És a dir, si $Z \sim N(0,1)$, aleshores

$$\lim_{n\to\infty} P(Z_n < z) = P(Z < z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

Observació. *En el context del teorema anterior, si es redefineix* Z_n *com*

$$Z_n = \frac{\sqrt{n}}{\sigma_n}(\mu_n - \mu) \qquad amb \qquad \sigma_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_n)^2,$$

aleshores la successió Z_n *també convergeix en llei a una normal de mitjana 0 i variància 1. Això és pràctic perquè permet utilitzar el teorema central del límit sense necessitat de conèixer la variància de* X*, és a dir* σ^2*.*

Aquest resultat permet calcular intervals de confiança per μ. Per fer-ho observem que

$$P\left(\mu \in (\mu_n - \varepsilon, \mu_n + \varepsilon)\right) = P\left(-\varepsilon < \mu_n - \mu < \varepsilon\right) = P\left(-\frac{\sqrt{n}}{\sigma_n}\varepsilon < \frac{\sqrt{n}}{\sigma_n}(\mu_n - \mu) < \frac{\sqrt{n}}{\sigma}\varepsilon\right)$$
$$\approx P\left(-\frac{\sqrt{n}}{\sigma_n}\varepsilon < Z < \frac{\sqrt{n}}{\sigma_n}\varepsilon\right).$$

Així, si $z(\alpha) > 0$ es defineix com el valor tal que $P(-z(\alpha) < Z < z(\alpha)) = 1 - \alpha$, es té que

$$\mu \in (\hat{\mu}_n - \varepsilon, \hat{\mu}_n + \varepsilon) \qquad \text{amb} \qquad \varepsilon = \frac{\hat{\sigma}_n z(\alpha)}{\sqrt{n}}$$

amb un grau de confiança de $1 - \alpha$, on

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{i} \qquad \hat{\sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2$$

i $x_1, x_2, \ldots, x_n$ són n realitzacions de la variable X.

Se solen utilitzar graus de confiança del 95% o del 99%, la qual cosa equival a prendre $\alpha = 0.05$ i $\alpha = 0.01$ respectivament, i per aquests valors d'α els valors de $z(\alpha)$ són $z(0.05) \approx 1.96$ i $z(0.01) \approx 2.58$ respectivament.

Observació 1. L'interval de confiança que s'obté és només aproximat (ja que es basa en un resultat assimptòtic, i per tant l'aproximació és tant millor com més gran sigui n). Si n no és prou gran pot passar, de fet, que $\hat{\sigma}_n = 0$ i es conclogui erròniament que μ és igual a μ_n amb un grau de confiança del 100%. Per exemple, si X és una variable aleatòria que val 0 amb probabilitat p i 1 amb probabilitat $1 - p$ i la probabilitat p és molt petita, aleshores si n és petit és probable que $\hat{\mu}_n = 1$ i $\hat{\sigma}_n = 0$.

Observació 2. Si es vol donar un interval de confiança en termes de la desviació típica σ, i no de l'estimador σ_n, caldrà conèixer una cota superior de σ (noteu que en casos pràctics difícilment es disposarà del valor exacte de σ, ja que si d'una variable aleatòria no en sabem la seva mitjana el més segur és que tampoc en sapiguem la seva desviació típica). Si la variable aleatòria X està acotada entre a i b, és a dir si els valors que pot prendre X estan inclosos a l'interval $[a, b]$, aleshores es té que $\sigma \leq (b - a)/2$. Per veure que això és així n'hi ha prou en adonar-se que una variable aleatòria que pren valors entre a i b no pot superar la dispersió d'una variable aleatòria que pren el valor a amb probabilitat $1/2$ i el valor b amb probabilitat $1/2$, i la desviació típica d'aquesta última variable és precisament $(b - a)/2$.

Exercici. Expliqueu com calcularíeu la integral

$$\int_0^2 \int_{-1}^1 x^2 + y^2 \, dxdy$$

utilitzant el mètode de Montecarlo. Especifiqueu una variable aleatòria l'esperança de la qual sigui el valor de la integral. Després escriviu un programa (en pseudocodi) que calculi el valor estimat de la integral i doni l'interval de confiança (aproximat) del 99% quan s'utilitzen n realitzacions de la variable aleatòria especificada.

2.2 Reducció de la variància

Com s'ha vist, l'error en el mètode de Montecarlo es comporta com $\sigma/\sqrt{n}$ quan n es fa gran, on σ^2 és la variància de la variable aleatòria de la qual es vol conèixer la mitjana. Hi ha dos maneres de reduir l'error per tant, o bé augmentar n (la qual cosa no té secret però comporta temps computacional que potser no es té) o bé treballar amb variables aleatòries alternatives que tinguin el mateix valor esperat però menor variància. A continuació veurem diverses estratègies per aconseguir això.

Per concretar l'exposició ens centrarem en el problema de calcular la integral

$$\int_0^1 g(x)dx,$$

que equival a dir que es vol trobar l'esperança $\mathbb{E}(Y)$ d'una variable aleatòria $Y = g(X)$, on X és una distribució uniforme a l'interval $(0, 1)$. Anem a veure com trobar variables $\tilde{Y}$ alternatives tals que $\mathbb{E}(\tilde{Y}) = \mathbb{E}(Y)$ però $\mathrm{Var}(\tilde{Y}) < \mathrm{Var}(Y)$, de manera que l'estimador

$$\frac{1}{n}\sum_{i=1}^n \tilde{Y}_i$$

estigui més concentrat que l'estimador

$$\frac{1}{n}\sum_{i=1}^n Y_i,$$

on $\tilde{Y}_i$ i Y_i són còpies independents de les variables aleatòries $\tilde{Y}$ i Y respectivament.

Mostreig d'importància

La idea del mostreig d'importància és reescriure la mitjana de $g(X)$ on X és una variable aletòria que té densitat f_X (anomenada *densitat nominal*) en termes d'una altra variable aleatòria $\tilde{X}$ que té densitat $f_{\tilde{X}}$ (anomenada *densitat d'importància*). Concretament es reescriu

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

com

$$\mathbb{E}\left(\frac{g(\tilde{X})f_X(\tilde{X})}{f_{\tilde{X}}(\tilde{X})}\right) = \int_{-\infty}^{\infty} \frac{g(x)f_X(x)}{f_{\tilde{X}}(x)} f_{\tilde{X}}(x)dx.$$

Aleshores, com que les integrals anteriors donen el mateix, es té que la variable aleatòria $Y := g(X)$ i la variable aleatòria

$$\tilde{Y} := \frac{g(\tilde{X})f_X(\tilde{X})}{f_{\tilde{X}}(\tilde{X})}$$

tenen la mateixa esperança. Per tant, si $\tilde{Y}$ té menor variància que Y, la mitjana empírica utilitzant n realitzacions de la variable $\tilde{Y}$ serà més precisa que la mitjana empírica utilitzant n realitzacions de la variable Y. És clar que per poder generar realitzacions de la variable $\tilde{Y}$ hem de ser capaços de generar realitzacions de la variable $\tilde{X}$ (com que la variable $Y = g(X)$ forma part del problema original es dona per fet que és possible generar realitzacions de la variable X).

Observem que la variància de $\tilde{Y}$, si denotem $\mu = \mathbb{E}(\tilde{Y})$, és

$$\begin{aligned}\mathrm{Var}(\tilde{Y}) = \mathbb{E}((\tilde{Y}-\mu)^2) &= \int_{-\infty}^{\infty} \left(\frac{g(x)f_X(x)}{f_{\tilde{X}}(x)} - \mu\right)^2 f_{\tilde{X}}(x)dx \\ &= \int_{-\infty}^{\infty} \left(\frac{g(x)f_X(x) - \mu f_{\tilde{X}}(x)}{f_{\tilde{X}}(x)}\right)^2 f_{\tilde{X}}(x)dx,\end{aligned}$$

la qual cosa indica que si $g(x)$ és positiva, aleshores la densitat d'importància òptima és

$$f_{\tilde{X}}(x) = \frac{g(x)f_X(x)}{\mu},$$

ja que té variància nul·la (la qual cosa vol dir que $\tilde{Y}$ pren el valor μ amb probabilitat 1, i per tant amb una única realització de $\tilde{Y}$, és a dir de $\tilde{X}$, seríem capaços de trobar l'esperança de Y amb una certesa absoluta). És clar, però, que aquesta densitat òptima no la podem conèixer ja que requereix conèixer el valor de μ que és precisament el que busquem. No obstant, ens dóna una idea de com han de ser les densitat d'importància útils: són aquelles que es troben a prop d'aquesta densitat òptima, i això equival a dir que la variable aleatòria $\tilde{X}$ ha de prendre valors més freqüentment en zones on la funció $g(x)f_X(x)$ pren valors alts, mentre que ha de prendre valors menys freqüentment en zones on la funció $g(x)f_X(x)$ pren valors propers a zero. Concretament s'ha d'aconseguir que $f_{\tilde{X}}(x)$ sigui tan proporcional a $g(x)f_X(x)$ com es pugui (en altres paraules, que la funció $g(x)f_X(x)/f_{\tilde{X}}(x)$ sigui tan constant com es pugui).

Si $g(x)$ no és positiva, aleshores es pot demostrar que la densitat òptima és

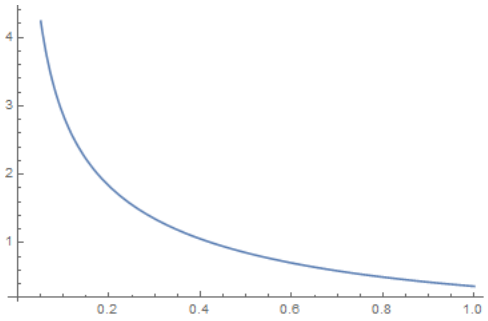
$$f_{\tilde{X}}(x) = \frac{|g(x)|f_X(x)}{\mu},$$

tot i que en aquest cas la variància de $\tilde{Y}$ ja no és nul·la quan s'utilitza aquesta densitat òptima. L'important és que ens dóna igualment una densitat de referència per buscar densitats d'importància útils (en aquest cas cal que $|g(x)|f_X(x)/f_{\tilde{X}}(x)$ sigui tan constant com es pugui).

Exemple. Considerem la integral

$$\int_0^1 x^{-\frac{1}{2}} e^{-x} dx,$$

que com sabem equival a trobar l'esperança de $Y = g(X)$ on $X \sim \text{Unif}(0,1)$ i $g(x) = x^{-1/2}e^{-x}$. Observem que la funció no és acotada en l'interval $(0,1)$, té una asímptota vertical a $x = 0$. La gràfica de la funció és



Si es vol utilitzar el mètode de Montecarlo directament, utilitzant com a densitat nominal la densitat de X, aleshores l'estimació que obtenim per la mitjana de Y (que equival a la integral anterior) serà força dolenta ja que la variància de Y és infinita, com es veu calculant el segon moment de Y:

$$\mathbb{E}(Y^2) = \int_0^1 x^{-1}e^{-2x}dx = \infty$$

ja que a prop de $x = 0$ l'integrand es comporta com la funció x^{-1} que no és integrable. Per solucionar això cal considerar una densitat d'importància que permeti mostrejar els punts propers a $x = 0$ molt més que els punts propers a $x = 1$.

Com que la part de la funció $g(x) = x^{-1/2}e^{-x}$ que dóna problemes és el factor $x^{-1/2}$ i sabem calcular la integral

$$\int_0^1 x^{-\frac{1}{2}}dx = 2,$$

podem agafar com a densitat d'importància una densitat proporcional a la funció $x^{-\frac{1}{2}}\mathbb{1}_{(0,1)}(x)$, concretament la densitat

$$f_{\tilde{X}}(x) = \frac{1}{2}x^{-\frac{1}{2}}\mathbb{1}_{(0,1)}(x).$$

Aleshores, si som capaços de generar realitzacions d'una variable aleatòria $\tilde{X}$ que tingui per densitat $f_{\tilde{X}}$, podrem generar realitzacions de

$$\tilde{Y} := \frac{g(\tilde{X})f_X(\tilde{X})}{f_{\tilde{X}}(\tilde{X})} = \frac{\tilde{X}^{-\frac{1}{2}}e^{-\tilde{X}}}{\frac{1}{2}\tilde{X}^{-\frac{1}{2}}} = 2e^{-\tilde{X}},$$

i l'estimació de la mitjana de $\tilde{Y}$ que obtindrem serà relativament bona en comparació a l'estimació utilitzant realitzacions de Y, ja que $\tilde{Y}$ sí té variància finita ja que aquesta variable està acotada (perquè $\tilde{X} \in (0,1)$).

La pregunta que cal abordar, per tant, és si és factible generar realitzacions de $\tilde{X}$. A la següent secció s'explica com fer-ho.

2.3 Generació de variables aleatòries

La teoria sobre generació de variables aleatòries es divideix en dos temes. D'una banda s'estudia com generar realitzacions d'una variable uniforme a l'interval $(0,1)$. Es tracta de mètodes deterministes que generen seqüències de nombres que es distribueixen de forma molt similar al que seria una seqüència de realitzacions independents d'una uniforme a l'interval $(0,1)$. Com que els nombres que generen aquests mètodes no són nombres aleatoris pròpiament dits, es diu que aquests mètodes generen nombres pseudoaleatoris que es comporten com una $\text{Unif}(0,1)$. Per veure'n alguns d'ells podeu consultar `https://www.math.arizona.edu/~tgk/mc/book_chap3.pdf`.

L'altra tema consisteix en donar mètodes que permetin generar realitzacions de variables aleatòries arbitràries utilitzant realitzacions de variables aleatòries uniformes a l'interval $(0,1)$. A continuació s'expliquen dos mètodes per aconseguir això: el **mètode de la inversa** i el **mètode d'acceptació rebuig**.

Mètode de la inversa

Un mètode eficient per generar variables aleatòries és el del mostreig de la distribució inversa (o mètode de la inversa), el qual peremet generar realitzacions d'una variable aleatòria X si coneixem la seva funció de distribució, és a dir si coneixem què val

$$F_X(x) := P(X \leq x).$$

Per fer-ho cal definir

$$F_X^{-1}(u) := \inf\{x \in \mathbb{R} \,|\, F_X(x) \geq u\}, \tag{2.4}$$

que és una espècie de funció inversa de F_X (és la inversa si F_X és invertible, però noteu que F_X^{-1} així definida té sentit també si F_X no és invertible). El mètode es basa en el següent resultat:

Teorema. *Sigui F_X la funció de distribució de la variable X. Sigui F_X^{-1} la funció inversa generalitzada de F_X definida a* (2.4)*. Sigui $U \sim Unif(0,1)$. Aleshores la variable aleatòria $F_X^{-1}(U)$ té la mateixa distribució que la variable X.*

Prova. Per demostrar el teorema cal veure que $P(F_X^{-1}(U) < x) = F_X(x)$. Observem que si la condició $F_X^{-1}(U) < x$ equival a la condició $U < F_X(x)$, aleshores ja ho tenim ja que

$$P(F_X^{-1}(U) \leq x) = P(U \leq F_X(x)) = F_X(x)$$

on la darrera igualtat es deu del fet que U és una uniforme en l'interval $(0,1)$ i $F_X(x) \in [0,1]$. Vegem, doncs, que efectivament $F_X^{-1}(U) \leq x$ si i només si $U < F_X(x)$.

Comencem suposant que $F_X^{-1}(U) \leq x$, és a dir que

$$x \geq \inf\{y \in \mathbb{R} \,|\, F_X(y) \geq U\},$$

i vegem que aleshores necessàriament $U \leq F_X(x)$. Per fer-ho obserevem que el conjunt $\{y \in \mathbb{R} \,|\, F_X(y) \geq U\}$ és de la forma $[c, \infty)$ on $c = \inf\{y \in \mathbb{R} \,|\, F_X(y) \geq U\}$. Això és així perquè F_X és

no-decreixent i contínua per la dreta (perquè és la funció de distribució d'una variable aleatòria). Per tant $x \geq c$ implica que $x \in [c, \infty) = \{y \in \mathbb{R} \,|\, F_X(y) \geq U\}$ i per tant $F_X(x) \geq U$ com volíem veure.

Suposem ara que $F_X(x) \geq U$ i vegem que això implica $F_X^{-1}(U) \leq x$. En efecte, si $F_X(x) \geq U$ aleshores $x \in \{y \in \mathbb{R} \,|\, F_X(y) \geq U\}$ i en particular x és igual o major a l'ínfim del conjunt $\{y \in \mathbb{R} \,|\, F_X(y) \geq U\}$, és a dir

$$x \geq \inf\{y \in \mathbb{R} \,|\, F_X(y) \geq U\} = F_X^{-1}(U).$$

□

Exemple. Per generar realitzacions d'una exponencial de paràmetre λ, considerem la seva funció de distribució, que és

$$F_{\text{Exp}(\lambda)}(x) = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x} \qquad \text{si } x \geq 0$$

i $F_{\text{Exp}(\lambda)}(x) = 0$ si $x < 0$. Com que $F_{\text{Exp}(\lambda)}$ és invertible en $[0, \infty)$, la seva inversa serà la inversa generalitzada del teorema anterior, per tant per trobar $F^{-1}_{\text{Exp}(\lambda)}(u)$ resolem l'equació $F_{\text{Exp}(\lambda)}(x) = u$ per la incògnita x:

$$\begin{aligned} 1 - e^{-\lambda x} = u \quad &\Rightarrow \quad 1 - u = e^{-\lambda x} \quad \Rightarrow \quad \log(1-u) = -\lambda x \\ &\Rightarrow \quad x = -\frac{1}{\lambda}\log(1-u). \end{aligned}$$

Per tant, la variable

$$X = -\frac{1}{\lambda}\log(1-U)$$

on $U \sim \text{Unif}(0,1)$ es distribueix com una exponencial de paràmetre λ.

Exercici. Considereu la densitat

$$f_{\tilde{X}}(x) = \frac{1}{2} x^{-\frac{1}{2}} \mathbb{1}_{(0,1)}(x).$$

donada a l'exemple de l'apartat (2.2).

a) Doneu una fòrmula que permeti generar realitzacions d'una variable $\tilde{X}$ que tingui $f_{\tilde{X}}$ per funció de densitat.

b) Apliqueu la fòrmula que heu donat per estimar la mitjana de la variable $\tilde{Y} = 2e^{-\tilde{X}}$ utilitzant 10000 realitzacions de $\tilde{Y}$. Doneu l'interval de confiança del 95% per la mitjana de $\tilde{Y}$. Comproveu si el valor de la integral

$$\int_0^1 x^{-\frac{1}{2}} e^{-x} dx \approx 1.49365$$

es troba dins de l'interval que heu donat (en principi hi hauria d'estar 95 de cada 100 vegades).

Mètode d'acceptació rebuig

El problema del mètode de la inversa és que cal disposar de la inversa de la funció de distribució F_X^{-1} de la variable aleatòria X que es vol generar. Això representa problemes tècnics importants que quan no es té una fórmula explícita per F_X^{-1}, la qual cosa pot fa que el mètode de la inversa perdi eficiència i sigui complicat d'implementar en aquests casos.

El mètode d'acceptació rebuig permet generar realitzacions d'una variable aleatòria X fent avaluacions només de la seva funció de densitat, és a dir de la funció $f_X(x)$. Per poder utilitzar aquest mètode, cal, però, saber generar realitzacions d'una variables aleatòria Y amb una funció de densitat f_Y que satisfaci $f_X(x) \leq m f_Y(x)$ per un $m > 0$ fixat i per a tot $x \in \mathbb{R}$.

Donades X, Y, f_X, f_Y i m, l'algorisme del mètode és el següent:

1. Generem una realització de Y, que denotem per x.
2. Generem una relaització de $U \sim \text{Unif}(0,1)$, que denotem per u.
3. Si
$$u < \frac{f_X(x)}{m f_Y(x)}$$
aleshores acceptem x com a realització de X. En cas contrari rebutgem x i tornem al pas 1.

Per veure que aquest algorisme funciona, és a dir que les realitzacions així generades es distribueixen amb la llei de X, n'hi ha prou en veure que el vector aleatori $(Y, U m f_Y(Y))$ es distribueix uniformement sobre l'àrea sota la gràfica de la funció $m f_Y$. Observeu que aquest fet implica que les realitzacions $(x, u m f_Y(x))$ de $(Y, U m f_Y(Y))$ que satisfan $u m f_Y(x) < f_X(x)$ també es distribueixen uniformement sobre l'àrea sota la gràfica de f_X, la qual cosa implica que la quantitat d'aquestes realitzacions amb la primera coordenada continguda a $[x, x + dx]$ és proporcional a $f_X(x)dx$ (si dx és arbitràriament petit), i per tant la densitat de la primera coordenada del vector aleatori (anomenada distribució marginal del vector aleatori respecte la primera component) és f_X.

La següent proposició mostra que, en efecte, el vector aleatori $(Y, U m f_Y(Y))$ es distribueix uniformement sobre l'àrea sota la gràfica de $m f_Y$.

Proposició. *La funció de densitat del vector aleatori $(Y, U m f_Y(Y))$ és*
$$f_{(Y, U m f_Y(Y))}(x, y) = \frac{1}{m} \mathbb{1}_{[0, m f_Y(x)]}(y),$$
és a dir la densitat és constant igual a $1/m$ si (x, y) satisfà $0 \leq y \leq m f_Y(x)$ (és a dir si el punt (x, y) està sobre l'àrea sota la gràfica de la funció $m f_Y$), i la densitat és 0 si el punt (x, y) no pertany a aquesta regió.

Prova. Reescrivim la densitat del vector aleatori com
$$f_{(Y, U m f_Y(Y))}(x, y) = \int_{-\infty}^{\infty} f_{(Y, U m f_Y(Y)) | Y = z}(x, y) f_Y(z) dz = \int_{-\infty}^{\infty} f_{(z, U m f_Y(z))}(x, y) f_Y(z) dz.$$

Observem ara que

$$f_{(z,Umf_Y(z))}(x,y) = \delta_z(x) f_{Umf_Y(z)}(y)$$

ja que la primera component del vector $(z, Umf_Y(z))$ no és aleatòria (és la constant z), i per tant la densitat respecte aquesta component està concentrada a z. La funció δ_z denota la delta de Dirac centrada en z, una funció que integra 1 i val 0 a tot arreu menys a z. Aquesta funció es pot interpretar com el límit de les densitats d'uniformes sobre intervals $[z-\varepsilon, z+\varepsilon]$ quan $\varepsilon \to 0$, és a dir com $\delta_z(x) = \lim_{\varepsilon\to 0} 1/(2\varepsilon)\mathbb{1}_{[z-\varepsilon,z+\varepsilon]}(x)$. Aleshores

$$\int_{-\infty}^{\infty} f_{(z,Umf_Y(z))}(x,y) f_Y(z) dz = \int_{-\infty}^{\infty} \delta_z(x) f_{Umf_Y(z)}(y) f_Y(z) dz = f_{Umf_Y(x)}(y) f_Y(x),$$

i com que $Umf_Y(x)$ és una uniforme sobre l'interval $[0, mf_Y(x)]$, es té

$$f_{Umf_Y(x)}(x) = \frac{1}{mf_Y(x)} \mathbb{1}_{[0,mf_Y(x)]}(y),$$

i per tant

$$f_{(Y,Umf_Y(Y))}(x,y) = \frac{1}{m} \mathbb{1}_{[0,mf_Y(x)]}(y). \qquad \square$$

Exemple. Suposem que volem generar punts uniformement distribuïts sobre l'el·lipse

$$S = \left\{ (x,y) \in \mathbb{R}^2 \,\middle|\, \frac{x^2}{4} + y^2 = 1 \right\}$$

Per fer-ho considereu la següent parametrització de S

$$\begin{array}{rccc} \Gamma: & [0,2\pi] & \to & \mathbb{R}^2 \\ & \alpha & \mapsto & (2\cos(\alpha), \sin(\alpha)) \end{array}.$$

La idea és definir una variable aleatòria X que prengui valors al segment $[0,2\pi]$ de manera que $\Gamma(X)$ serà un vector aleatòri amb valors sobre S. La pregunta és quina ha de ser la densitat de la variable X per tal que els punts sobre S estiguin distribuïts uniformement. Per respondre aquesta pregunta observem que la velocitat en que es recorre l'el·lipse amb aquesta parametrització ve donada per la derivada de Γ, concretament $\|\Gamma'(\alpha)\|$ dona com de ràpid s'està parametritzant l'el·lipse en el punt $\Gamma(\alpha)$. Com que la densitat de punts sobre l'el·lipse volem que sigui uniforme, cal imposar que la densitat de X en α sigui proporcional a $\|\Gamma'(\alpha)\|$ (ja que el segment $[\alpha, \alpha+d\alpha]$ es transforma en un arc d'el·lipse de longitud $\|\Gamma'(\alpha)\|d\alpha$ quan $d\alpha$ és arbitràriament petit, i per tant la quantitat de punts que hi ha d'haver entre $[\alpha, \alpha+d\alpha]$ ha de ser proporcional a $\|\Gamma'(\alpha)\|d\alpha$), és a dir cal que

$$f_X(\alpha) = \frac{\|\Gamma'(\alpha)\|}{c} \qquad \text{amb} \quad c = \int_0^{2\pi} \|\Gamma'(a)\| da.$$

Ara bé, com que $\|\Gamma'(\alpha)\| = \sqrt{4\sin(\alpha)^2 + \cos(\alpha)^2} \leq \sqrt{5}$, es té

$$f_X(\alpha) = \frac{\|\Gamma'(\alpha)\|}{c} \leq \frac{\sqrt{5}}{c}$$

i per tant per generar realitzacions de X es pot utilitzar el mètode d'acceptació rebuig amb

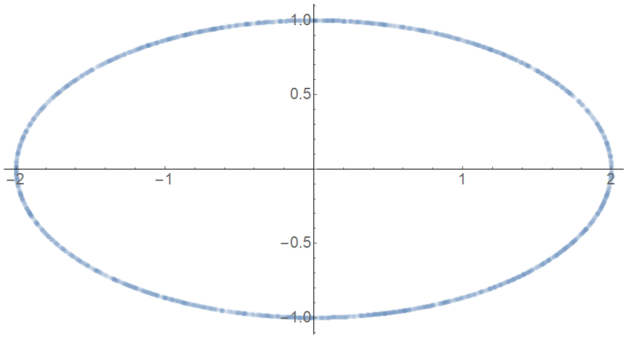
$$Y \sim \text{Unif}(0,2\pi) \qquad \text{i} \qquad m = 2\pi\frac{\sqrt{5}}{c}.$$

Observem que per aplicar el mètode d'acceptació rebuig amb aquesta m no cal conèixer el valor de c, ja que el quocient que s'utilitza al punt 3. de l'algorisme esdevé

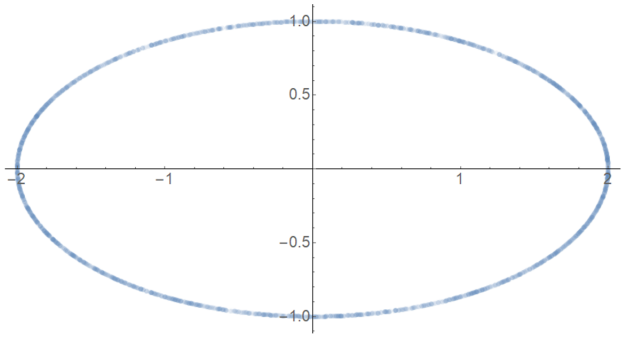
$$\frac{f_X(\alpha)}{m f_Y(\alpha)} = \frac{\frac{\|\Gamma'(\alpha)\|}{c}}{2\pi \frac{\sqrt{5}}{c} \frac{1}{2\pi}} = \frac{\|\Gamma'(\alpha)\|}{\sqrt{5}} = \frac{\sqrt{4\sin(\alpha)^2 + \cos(\alpha)^2}}{\sqrt{5}}.$$

A la figura següent es mostra la distribució de punts sobre l'el·lipse que s'obté quan s'utilitza l'algorisme d'acceptació rebuig per a generar realitzacions de X juntament amb la distribució de punts quan s'agafen angles uniformement distribuïts sobre l'interval $[0, 2\pi]$. Observeu que en aquest segon cas els punts tendeixen a estar més acumulats en zones de l'el·lipse properes als punts $(-2, 0)$ i $(2, 0)$ i més dispersos en zones properes als punts $(0, -1)$ i $(0, 1)$ (la diferència seria més evident si la diferència entre els eixos de l'el·lipse fos més gran).

```
l = {};
For[i = 1, i < 3000, i++,
 While[True,
  y = RandomReal[2 Pi];
  u = RandomReal[1];
  If[u < Sqrt[4 Sin[y]^2 + Cos[y]^2] / Sqrt[5],
   Break[]
  ];
 ];
 l = Append[l, {2 Cos[y], Sin[y]}]
]
ListPlot[l]
```




```
l = {};
For[i = 1, i < 3000, i++,
 y = RandomReal[2 Pi];
 l = Append[l, {2 Cos[y], Sin[y]}]
]
ListPlot[l]
```



Problemes per practicar

1. Considereu el conjunt C donat a l'Exemple 1. Expliqueu com aplicaríeu un mètode de Montecarlo per calcular la integral

$$\int_C g(x, y) dx dy$$

on g és una funció integrable sobre C que pren valors entre a i b ($a < b$). Especifiqueu com donaríeu un interval de confiança del 95% utilitzant una cota superior per la desviació típica de la variable aleatòria que utilitzeu per calcular la integral.

2. Considereu la integral

$$\int_0^{10} \left(1 + \frac{\sin(x^2)}{10}\right) e^{-x} dx.$$

a. Expliqueu com calcularíeu la integral anterior utilitzant realitzacions d'una uniforme.

b. Expliqueu com calcularíeu la integral anterior utilitzant realitzacions d'una variable aleatòria X amb funció de densitat

$$f_X(x) = ce^{-x}\mathbb{1}_{[0,10]}(x)$$

on la constant c és tal que $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

c. Quin dels dos mètodes espereu que aproximi millor la integral?

3. Expliqueu com generaríeu realitzacions de la variable X definida a l'apartat b. del problema anterior utilitzant realitzacions d'una uniforme sobre l'interval $[0, 1]$.

4. Exercici proposat a la pàgina 12 d'aquest document.

5. Escriviu un codi (o pseudocodi) per una funció que retorni una realització d'una variable aleatòria X la funció de distribució de la qual és

$$F_X(x) = \begin{cases} 0 & \text{si} \quad x < 1 \\ (x-1)/2 & \text{si} \quad 1 \leq x < 2 \\ \left(3 + \frac{x-2}{1+(x-2)}\right)/4 & \text{si} \quad 2 \leq x \end{cases}$$

6. Expliqueu com generaríeu punts aleatoris uniformement distribuïts sobre el tros de paràbola

$$S = \{(x, y) \in \mathbb{R}^2 \mid y = 4x^2 \quad \text{amb} \quad x \in [-1, 1]\}$$

Capítol 3

Mètodes numèrics per resoldre EDOs

En aquest tema considerem el problema de valor inicial general de la forma

$$\begin{cases} y'(t) = f(y(t), t) \\ y(t_0) = y_0 \end{cases}$$

amb f una funció contínua i Lipschitz respecte el primer argument (la variable y), la qual cosa assegura que existeix una única funció solució del problema (de manera que té sentit plantejar-se trobar la solució ja sigui analíticament o numèricament).

La solució del problema anterior no sempre es pot calcular utilitzant tècniques analítiques com poden ser la *separació de variables*, l'ús de *factors integrants* o fent *canvis de variables*, per citar-ne alguns.

Per la majoria de casos pràctics cal implementar mètodes numèrics per aproximar la solució del problema anterior. En aquests casos el que es vol conèixer és el valor de la funció y en un punt t_f donat (f de final). Com que t_f pot estar molt allunyat de t_0, no és raonable pensar que es pugui aproximar $y(t_f)$ amb una única operació. Sí ens podem plantejar, però, aproximar $y(t_0 + h)$ si h és petit, ja que tenim informació sobre què val y en t_0 (la condició inicial $y(t_0) = y_0$) i sobre com varia la funció y (l'equació diferencial $y'(t) = f(y(t), t)$). La idea és, per tant, aproximar $y(t_f)$ fent aproximacions prèvies de y en els punts $t_1 = t_0 + h$, $t_2 = t_0 + 2h$, $t_3 = t_0 + 3h$, etc., de manera que per aproximar y en el punt $t_{i+1} = t_0 + (i+1)h$ s'utilitza l'aproximació de y en el punt $t_i = t_0 + ih$.

3.1 Alguns esquemes a mode introductori

Esquemes de Taylor

Vegem un primer exemple que il·lustra aquesta mecànica. Observeu que, fixat t, per aproximar $y(t + h)$ en termes de $y(t)$ i t podem utilitzar el desenvolupament de Taylor

$$y(t + h) = y(t) + hy'(t) + O(h^2)$$

on $O(h^p)$ per $p > 0$ es pot interpretar com una funció que satisfà $\limsup_{h\to 0} |O(h^p)|/h^p < \infty$, on $\limsup$ vol dir limit superior. En paraules això equival a dir que $O(h^p)/h^p$ està acotat quan h tendeix a 0. Observeu que de la definició se segueix que $\lim_{h\to 0} O(h^p)/h^q = 0$ si $0 \leq q < p$.

Ara bé, $y'(t) = f(y(t), t)$, per tant si h és petit es pot aproximar $y(t+h)$ com

$$y(t+h) \approx y(t) + y'(t)h = y(t) + hf(y(t), t).$$

Si denotem per Y_i l'aproximació de $y(t_i)$ on $t_i = t_0 + ih$, aleshores l'aproximació anterior dóna lloc a l'esquema numèric

$$\begin{cases} Y_{i+1} = Y_i + hf(Y_i, t_i) \\ Y_0 = y_0. \end{cases}$$

Aquest esquema numèric es coneix com a **Mètode d'Euler explícit**. És un esquema **monopàs**, que vol dir que per calcular Y_{i+1} (l'aproximació de y en el punt t_{i+1}) només s'utilitza Y_i (l'aproximació de y calculada en el pas previ, és a dir en el punt t_i). Després veurem un exemple d'esquema **multipàs** en el que per calcular Y_{i+1} s'utilitzen les aproximacions de y en múltiples passo previs (per exemple en un esquema de dos passos es calcula Y_{i+1} utilitzant Y_i i Y_{i-1}). D'altra banda és un esquema **explícit** perquè Y_{i+1} s'expressa de forma explícita en termes de Y_i. Al final de la secció veurem el mètode d'Euler implícit que és un esquema **implícit** ja que Y_{i+1} s'expressarà en funció de Y_i de forma implícita.

El mètode d'Euler és un cas particular d'**esquema de Taylor**, concretament l'esquema de Taylor d'ordre 1. Si considerem més termes del desenvolupament de Taylor de la funció $y(t+h)$ respecte h, podem obtenir esquemes de Taylor d'ordre superior. Per exemple, l'esquema de Taylor d'ordre 2 s'obté observant que

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + O(h^3),$$

de manera que si h és petit es pot aproximar $y(t+h)$ com

$$y(t+h) \approx y(t) + hy'(t) + \frac{h^2}{2}y''(t)$$

on observeu que $y'(t)$ i $y''(t)$ es poden escriure en termes de t i $y(t)$ gràcies a l'equació diferencial. En efecte, $y'(t) = f(y(t), t)$ i

$$y''(t) = \partial_1 f(y(t), t)y'(t) + \partial_2 f(y(t), t) = \partial_1 f(y(t), t)f(y(t), t) + \partial_2 f(y(t), t).$$

Així, l'aproximació anterior dóna lloc a l'esquema de Taylor d'ordre 2:

$$\begin{cases} Y_{i+1} = Y_i + hf(Y_i, t_i) + \frac{h^2}{2}\left(\partial_1 f(Y_i, t_i)f(Y_i, t_i) + \partial_2 f(Y_i, t_i)\right) \\ Y_0 = y_0. \end{cases}$$

Observació. Els esquemes de Taylor d'ordre més gran que 1 requereixen conèixer les derivades de la funció f, i això fa que no sigui fàcil automatitzar la seva implementació en problemes generals. Són útils, però, quan es vol estudiar una equació diferencial particular i acotar de forma precisa l'error que es comet.

Esquemes d'Adams (un exemple d'esquema multipàs)

Integrant a banda i banda l'equació $y'(t) = f(y(t), t)$ entre t i $t + h$ s'obté de la relació

$$y(t + h) = y(t) + \int_t^{t+h} f(y(\tau), \tau) d\tau \tag{3.1}$$

Per aproximar la funció $\tau \mapsto f(y(\tau), \tau)$ utilitzant només avaluacions de la funció y sobre valors de τ menors o iguals a t (que se suposa que són valors de y que hem aproximat en passos anteriors) podem utilitzar, per exemple la recta que passa pels punts $(t-h, f(y(t-h), t-h))$ i $(t, f(y(t), t))$, que és la recta

$$r(\tau) = \frac{f(y(t), t) - f(y(t-h), t-h)}{h}(\tau - t) + f(y(t), t).$$

Això ens dona l'aproximació

$$y(t+h) = y(t) + \int_t^{t+h} f(y(\tau), \tau) d\tau \approx y(t) + \int_t^{t+h} r(\tau) d\tau = y(t) + h\frac{3f(y(t), t) - f(y(t-h), t-h)}{2}.$$

L'aproximació anterior es correspon amb un esquema de dos passos (perquè per aproximar $y(t+h)$ s'utilitza $y(t)$ i $y(t-h)$). Concretament l'esquema s'escriu en termes de Y_i (on, com abans, Y_i denota l'aproximació de $y(t_i)$ amb $t_i = t_0 + ih$) com

$$\begin{cases} Y_{i+1} = Y_n + \frac{h}{2}(3f(Y_i, t_i) - f(Y_{i-1}, t_{i-1})) \\ Y_0 = y_0 \\ Y_1 \text{ es calcula amb un mètode monopàs a partir de } Y_0. \end{cases}.$$

Si enlloc d'utilitzar la recta que passa pels punts $(t-h, f(y(t-h), t-h))$ i $(t, f(y(t), t))$ haguèssim aproximat la funció $\tau \mapsto f(y(\tau), \tau)$ utilitzant la paràbola que passa pels punts $(t-2h, f(y(t-2h), t-2h))$, $(t-h, f(y(t-h), t-h))$ i $(t, f(y(t), t))$, obtindríem un esquema de 3 passos (en principi més precís que el de dos passos perquè la funció estaria millor aproximada si h és petit). En general, si s'utilitza el polinomi interpolador sobre els punts $(t-kh, f(y(t-kh), t-kh)), \ldots, (t-h, f(y(t-h), t-h)), (t, f(y(t), t))$ per aproximar la funció $\tau \mapsto f(y(\tau), \tau)$ s'obté un mètode de $k+1$ passos (concretament el mètode d'Adams de $k+1$ passos).

Un exemple d'esquema de Runge-Kutta

Una altra manera d'aproximar la funció $\tau \mapsto f(y(\tau), \tau)$ entre t i $t + h$ és utilitzant la recta que passa pels punts $(t, f(y(t), t))$ i $(t+h, f(y(t+h), t+h))$. El problema que té aquesta aproximació és que no coneixem què val $y(t+h)$ (això és precisament el que volem calcular), i per tant no podem utilitzar l'equació d'aquesta recta. Observeu, però, que el punt $(t+h, f(y(t+h), t+h))$ és proper al punt $(t+h, f(y(t) + hf(y(t), t), t+h))$ si h és petit, ja que $y(t+h) = y(t) + hf(y(t), t) + o(h)$. Per tant, podem aproximar la funció $\tau \mapsto f(y(\tau), \tau)$ entre t i $t+h$ utilitzant la recta que passa pels punts $(t, f(y(t), t))$ i $(t+h, f(y(t) + hf(y(t), t), t+h))$. Com que la integral d'aquesta recta entre t i $t+h$ és l'àrea del trapezi que formen els punts $(t, 0)$,$(t, f(y(t), t))$,$(t+h, f(y(t) + hf(y(t), t), t+h))$ i $(t+h, 0)$, i aquesta àrea val (utilitzant la fórmula de l'àrea del trapezi)

$$h\frac{f(y(t), t) + f(y(t) + hf(y(t), t), t+h)}{2},$$

es té l'aproximació

$$y(t+h) = y(t) + \int_t^{t+h} f(y(\tau),\tau) \approx y(t) + h\frac{f(y(t),t) + f(y(t)+hf(y(t),t),t+h)}{2},$$

que dona l'esquema d'un pas

$$\begin{cases} Y_{i+1} = Y_i + \frac{h}{2}(f(Y_i,t_i) + f(Y_i + hf(Y_i,t_i),t_{i+1})) \\ Y_0 = y_0 \end{cases},$$

que és un exemple de la família d'esquema de Runge-Kutta, la qual es descriurà més endavant.

Un exemple d'esquema implícit (Mètode d'Euler implícit)

Si s'escriu el desenvolupament en sèrie de potències de $y(t)$ centrat en el punt $t+h$ es té

$$y(t) = y(t+h) - y'(t+h)h + \frac{1}{2}y''(t+h)h^2 - \frac{1}{3!}y'''(t+h)h^3 + \dots$$

Utilitzant només els dos primer termes del desenvolupament anterior es té l'aproximació

$$y(t) \approx y(t+h) - y'(t+h)h = y(t+h) - hf(y(t+h),t+h),$$

que equival a $y(t+h) \approx y(t) + hf(y(t+h),t+h)$, la qual cosa dona l'esquema numèric

$$\begin{cases} Y_{i+1} = Y_i + hf(Y_{i+1},t_{i+1}) \\ Y_0 = y_0 \end{cases},$$

que és un esquema implícit perquè el valor de Y_{i+1} no s'expressa de forma explícita en termes de Y_i, i de fet per calcular Y_{i+1} caldria utilitzar algun mètode de cerca de solucions d'equacions no lineals (per exemple el mètode de Newton). Tot i que els esquemes implícits poden semblar poc pràctics, veurem que presenten algunes propietats que els fan adients en certes situacions.

3.2 Esquemes monopàs generals: convergència i consistència

Un esquema monopàs general es pot expressar en la forma

$$Y_{i+1} = Y_i + h\phi(t_i, Y_i, h; f)$$

on ϕ és una funció que depèn del node t_i, el valor de l'aproximació al pas i (és a dir Y_i), el pas h i la funció f (la dependència de la funció f inclou que es puguin prendre derivades de f o avaluar f en qualssevol punt).

Noteu que en el cas de l'esquema d'Euler explícit es té

$$\phi(t_i, Y_i, h; f) = f(Y_i, t_i)$$

i per l'esquema de Taylor d'ordre 2 es té

$$\phi(t_i, Y_i, h; f) = f(Y_i, t_i) + \frac{h}{2}\left(\partial_1 f(Y_i, t_i) f(Y_i, t_i) + \partial_2 f(Y_i, t_i)\right).$$

Com hem comentat, el que es vol d'un esquema numèric és que ens permeti aproximar bé la solució d'un problema de valor inicial per a un valor t_f donat. Concretament cal que l'aproximació sigui tant millor com més petit es faci el pas h de l'esquema, i en particular cal que quan el pas tendeix a zero la solució numèrica convergeixi a la solució del problema de valor inicial. Recollim aquesta característica en la següent definició.

Definició. Es diu que un esquema

$$Y_{i+1} = Y_i + h\phi(t_i, Y_i, h; f)$$

amb condició inicial $Y_0 = y_0$ **convergeix** a la solució del PVI

$$\begin{cases} y'(t) = f(y(t), t) \\ y(t_0) = y_0 \end{cases}$$

si per tot t_f del domini de definició de la funció incògnita y es té que

$$\lim_{n\to\infty} Y_n(h_n) = y(t_f)$$

on $Y_n(h_n)$ és la solució al pas n de l'esquema quan s'utilitza el pas $h_n = (t_f - t_0)/n$.

Per analitzar la convergència d'un esquema és oportú considerar quin és l'error que es produeix quan s'aplica l'esquema a la solució exacte de l'equació diferencial $y'(t) = f(y(t), t)$. És a dir, considerem l'error $\text{error}(t)$ que fa certa la fórmula

$$y(t + h) = y(t) + h\phi(t, y(t), h; f) + \text{error}(t)$$

per un t donat. La quantitat que ens interessa de cara a estudiar la convergència és aquest error normalitzat pel pas h. El motiu d'aquesta normalització es deu al fet que si es divideix el pas per la meitat caldrà doblar el nombre de passos per arribar a t_f, la qual cosa implica que s'acumularan el doble d'errors. Per tant, l'error efectiu que es produeix en un pas de longitud h quan s'està al punt t és de l'ordre $\text{error}(t)/h$, ja que l'error es reparteix al llarg de tot el segment que s'avança. Aquest error normalitzat es denomina **error de truncament local**, i queda recollit en la següent definició.

Definició. L'error de truncament local en un punt t s'obté fent la diferència entre la solució de l'equació diferencial $y'(t) = f(y(t), t)$ al punt $y(t + h)$ i la solució que dóna l'esquema partint de $y(t)$ i utilitzant el pas h, i dividint aquesta diferència per h. Es denota per $\tau(h, t)$ (ja que depèn tant del punt t com del pas h utilitzat), i formalment s'expressa com

$$\tau(h, t) = \frac{y(t + h) - y(t) - h\phi(t, y(t), h; f)}{h}.$$

Definició. Es diu que un esquema és d'**ordre** p (o que té ordre de consistència p) si p és el nombre natural més gran pel qual

$$\lim_{h\to 0} \frac{\tau(h,t)}{h^p} < \infty$$

per tot t. Es diu que un esquema és **consistent** si és d'ordre més gran o igual a 1.

Calcular l'ordre de consistència d'un mètode és relativament senzill, només cal expressar l'error de truncament (és a dir $\tau(h,t)$) com a sèrie de potències de h.

Exemple. L'esquema d'Euler explícit és d'ordre 1. En efecte, com que

$$y(t+h) = y(t) + y'(t)h + y''(t)\frac{h^2}{2} + O(h^3),$$

i $y'(t) = f(y(t),t)$, aleshores

$$\begin{aligned}\tau(h,t) &= \frac{y(t+h) - y(t) - hf(y(t),t)}{h} \\ &= \frac{y(t) + f(y(t),t)h + y''(t)\frac{h^2}{2} + O(h^3) - y(t) - hf(y(t),t)}{h} = y''(t)\frac{h}{2} + \frac{o(h^3)}{h}.\end{aligned}$$

Per tant

$$\lim_{h\to 0} \frac{\tau(h,t)}{h} = \lim_{h\to 0} \frac{y''(t)\frac{h}{2} + \frac{O(h^3)}{h}}{h} = \frac{y''(t)}{2} + \lim_{h\to 0} \frac{O(h^3)}{h^2} = \frac{y''(t)}{2} < \infty,$$

si la derivada segona de y està acotada en intervals tancats. Noteu que hem utilitzat que $\lim_{h\to 0} O(h^3)/h^2 = 0$, que se segueix del resultat general $\lim_{h\to 0} O(h^p)/h^q = 0$ si $0 \le q < p$ comentat abans.

Observeu d'altra banda que l'esquema no és d'ordre 2 perquè en general $y''(t)$ no és igual a zero per tot t, de manera que en aquests casos

$$\lim_{h\to 0} \frac{\tau(h,t)}{h^2} = \lim_{h\to 0} \frac{y''(t)\frac{h}{2} + \frac{O(h^3)}{h}}{h^2} = \lim_{h\to 0} \frac{y''(t)}{2h} + \frac{O(h^3)}{h^3} = \pm\infty.$$

ja que el terme $O(h^3)/h^3$ està acotat quan h tendeix a 0 per definició.

A continuació donem un resultat que implica que un esquema monopàs és convergent si és consistent. Això, permet determinar la convergència d'un mètode monopàs comprovant que l'esquema és consistent, que com hem dit equival a verificar que

$$\lim_{h\to 0} \frac{\tau(h,t)}{h} < \infty.$$

Veurem que en mètodes multipàs aquesta relació entre convergència i consistència ja no es té, i caldrà analitzar una altra característica de l'esquema per poder assegurar la seva convergència.

Teorema. *Suposem que $\phi(t,y,h;f)$ és contínua respecte els tres primers arguments i que a més a més és Lipschitz respecte el segon argument, és a dir suposem que existeix k tal que*

$$|\phi(t,y_1,h;f) - \phi(t,y_2,h;f)| \le k|y_1 - y_2|$$

per tota parella de punts y_1 i y_2. Aleshores (utilitzant la mateixa notació que en la definició de convergència) es té

$$|Y_n(h_n) - y(t_f)| \leq e^{k|t_f - t_0|}\left(e_0 + \frac{\tau(h_n)}{k}\right)$$

on $e_0 = Y_0 - y_0$ és l'error en la condició inicial (si sabem exactament quina és la condició inicial i es pot guardar exactament a la computadora, aleshores $e_0 = 0$) i on

$$\tau(h_n) := \max_{t\in[t_0,t_f]} |\tau(h_n, t)|.$$

Prova. Definim $e_i = Y_i - y(t_i)$ on $t_i = t_0 + ih_n$. La idea és acotar l'error en el pas $i+1$, és a dir $|e_{i+1}|$ en termes dels passos anteriors. Observem que (utilitzant la definició d'error local de truncament)

$$y(t_{i+1}) = y(t_i + h_n) = y(t_i) + h_n\phi(t_i, y(t_i), h_n; f) + h_n\tau(h_n, t_i)$$

i

$$Y_{i+1} = Y_i + h_n\phi(t_i, Y_i, h_n; f).$$

Per tant, utilitzant la desigualtat triangular i la propietat Lipschitz de ϕ respecte y i utilitzant h enlloc de h_n per simplificar la lectura, es té

$$\begin{aligned}|e_{i+1}| &\leq |Y_i + y(t_i)| + h|\phi(t_i, Y_i, h; f) - \phi(t_i, y(t_i), h; f)| + h|\tau(h, t_i)| \\ &\leq |e_i| + hk|e_i| + h\tau(h) = (1 + hk)|e_i| + h\tau(h).\end{aligned}$$

Aplicant la desigualtat anterior de forma recursiva s'obté que

$$|e_{i+1}| \leq |e_0|(1+hk)^{i+1} + h\tau(h)\sum_{j=0}^{i}(1+hk)^j = |e_0|(1+hk)^{i+1} + h\tau(h)\frac{(1+hk)^{i+1} - 1}{hk}.$$

Ara bé, com que $(1+hk)^{i+1} \leq e^{(i+1)hk}$ (això es conseqüència del fet que $1 + z \leq e^z$ per tot $z \geq 0$, i per tant també $(1+z)^{i+1} \leq e^{z(i+1)}$, que és el que es diu amb $z = hk$), la desigualtat anterior es pot reescriure com

$$\begin{aligned}|e_{i+1}| &\leq |e_0|(1+hk)^{i+1} + h\tau(h)\frac{(1+hk)^{i+1} - 1}{hk} \\ &= \left(|e_0| + \frac{\tau(h)}{k}\right)(1+hk)^{i+1} - \frac{\tau(h)}{k} \leq \left(|e_0| + \frac{\tau(h)}{k}\right)(1+hk)^{i+1} \\ &\leq \left(|e_0| + \frac{\tau(h)}{k}\right)e^{(i+1)hk}\end{aligned}$$

i en particular (utilitzant que $h = h_n = (t_f - t_0)/n$) es conclou

$$|Y_n(h_n) - y(t_f)| = |e_n| \leq \left(|e_0| + \frac{\tau(h_n)}{k}\right)e^{nh_nk} = \left(|e_0| + \frac{\tau(h_n)}{k}\right)e^{(t_f - t_0)k}$$

□

Observació. Si un esquema té ordre de consistència p, aleshores es té que

$$\tau(h) := \max_{t\in[t_0,t_f]} \tau(h,t) = O(h^p).$$

En particular es té $\tau(h) < Ch^p$ si h és petit per una constant C fixada. D'aquest fet se segueix que, si $e_0 = 0$, aleshores la velocitat de convergència de l'esquema a la solució $y(t_f)$ serà de l'ordre h^p. En altres paraules, si es divideix el pas h per 2, aleshores es dividirà per 2^p l'error de l'aproximació de la solució. En efecte:

$$|Y_n(h_n) - y(t_f)| \leq \frac{\tau(h_n)}{k} e^{(t_f-t_0)k} \leq \frac{C}{k} h_n^p e^{(t_f-t_0)k}.$$

3.3 Esquemes Implícits

Els mètodes monopàs implícits són esquemes que es poden escriure de la forma

$$Y_{i+1} = Y_i + h\tilde{\phi}(t_i, Y_i, Y_{i+1}, h; f). \tag{3.2}$$

Noteu que es diu implícit perquè per calcular Y_{i+1} no es té una fórmula explícita sino que s'ha de resoldre una equació implícita.

Observació. Tot i que per analitzar l'error de truncament local dels esquemes implícits va bé utilitzar la fórmula (3.2), tot esquema explícit es pot escriure com esquemes de la forma $Y_{i+1} = Y_i + h\phi(t_i, Y_i, h; f)$. En efecte, n'hi ha prou en reescriure la fòrmula (3.2) com

$$Y_{i+1} = Y_i + h\tilde{\phi}(t_i, Y_i, Y_i + hK(Y_i, t_i), h; f)$$

amb $K(Y_i, t_i)$ definit com la solució de l'equació $K(Y_i, t_i) = \tilde{\phi}(t_i, Y_i, Y_i + hK(Y_i, t_i), h; f)$. Per tant, els resultats donats a l'apartat 3.2 són també vàlids pels esquemes implícits.

Exemple. Considerem l'esquema d'Euler implícit, donat per

$$Y_{i+1} = Y_i + hf(Y_{i+1}, t_{i+1})$$

on recordeu que $t_{i+1} = t_i + h$. Anem a calcular l'ordre de consistència d'aquest mètode. Per fer-ho escrivim (com hem fet en l'exemple anterior) l'error de truncament local en sèrie de potències. L'error de truncament local d'aquest esquema és

$$\tau(h,t) = \frac{y(t+h) - y(t) - hf(y(t+h), t+h)}{h}.$$

Com que $f(y(t+h), t+h) = y'(t+h)$ és convenient reescriure $\tau(h,t)$ en termes de y' enlloc de f (ja que si no caldria derivar respecte la primera i la segona component de f i ens donarien expressions més llargues, tot i que el resultat sería igualment vàlid). Així tenim

$$\tau(h,t) = \frac{y(t+h) - y(t) - hy'(t+h)}{h}.$$

Ara, com que

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + O(h^3)$$

i

$$y'(t+h) = y'(t) + hy''(t) + O(h^2),$$

es té que

$$\tau(h,t) = \frac{y(t) + hy'(t) + \frac{h^2}{2}y''(t) + O(h^3) - y(t) - h\left(y'(t) + hy''(t) + O(h^2)\right)}{h}$$
$$\frac{-\frac{h^2}{2}y''(t) + O(h^3}{h}.$$

Per tant l'esquema és d'ordre 1 ja que

$$\lim_{h\to 0}\frac{\tau(h,t)}{h} = -\frac{1}{2}y''(t)$$

i no és d'ordre 2 ja que en general $y''(t)$ no és 0.

Exercici. Considereu la família d'esquemes

$$Y_{i+1} = Y_i + h(\alpha f(Y_i, t_i) + (1-\alpha)f(Y_{i+1}, t_{i+1}).$$

1. Proveu que l'esquema és consistent per tot α.
2. Existeix algun valor d'α pel qual l'esquema sigui d'ordre 2?.

3.4 Esquemes de Runge-Kutta

La idea dels esquemes RK és utilitzar una mitjana ponderada del camp $f(y(t), t)$ en punts "propers" a la corba $t \mapsto y(t)$ entre t_i i t_{i+h}.

Vegem un exemple senzill d'esquema RK. Considerem l'esquema

$$Y_{i+1} = Y_i + h(c_1K_1(Y_i, t_i) + c_2K_2(Y_i, t_i))$$

amb

$$K_1(Y_i, t_i) = f(Y_i, t_i)$$
$$K_2(Y_i, t_i) = f(Y_i + hb_{21}K_1(Y_i, t_i), t_i + ha_2)$$

on c_1, c_2, b_{21} i a_2 són paràmetres que cal ajustar perquè l'esquema tinguí ordre de consistència el més gran possible. Per determinar el valor dels paràmetres, cal, per tant, escriure l'error de truncament local en sèrie de potències del pas h. Per fer-ho recordem que

$$h\tau(h,t) = y(t+h) - y(t) - h(c_1K_1(y(t), t) + c_2K_2(y(t), t)),$$

i observem que

$$y(t+h) = y(t) + y'(t)h + y''(t)h^2/2 + O(h^3),$$

que $K_1(y(t),t) = y'(t)$ i

$$K_2(y(t),t) = f(y(t),t) + \partial_1 f(y(t),t)y'(t)hb_{21} + \partial_2 f(y(t),t)ha_2 + O(h^2).$$

Per tant, utilitzant que $y''(t) = \partial_1 f(y(t),t)y'(t) + \partial_2 f(y(t),t)$, es té

$$\begin{aligned} h\tau(h,t) &= y(t) + y'(t)h + y''(t)h^2/2 + O(h^3) - y(t) - hc_1y'(t) \\ &\quad - hc_2(y'(t) + \partial_1 f(y(t),t)y'(t)hb_{21} + \partial_2 f(y(t),t)ha_2 + O(h^2)) \\ &= y'(t)h(1 - c_1 - c_2) \\ &\quad + \partial_1 f(y(t),t)y'(t)h^2\left(\frac{1}{2} - c_2b_{21}\right) + \partial_2 f(y(t),t)y'(t)h^2\left(\frac{1}{2} - c_2a_2\right) + O(h^3) \end{aligned} \tag{3.3}$$

de manera que l'esquema serà d'ordre 2 si

$$\begin{cases} 0 = 1 - c_1 - c_2 \\ 0 = \dfrac{1}{2} - c_2b_{21} \\ 0 = \dfrac{1}{2} - c_2a_2 \end{cases}.$$

Això és un sistema de tres equacions amb 4 incògnites. Hi ha infinites solucions que es poden expressar en termes d'un paràmetre $\alpha \neq 0$ com

$$c_2 = \alpha, \quad c_1 = 1 - \alpha, \quad a_2 = \frac{1}{2\alpha} \quad \text{i} \quad b_{21} = \frac{1}{2\alpha}.$$

Per cada valor d'$\alpha \neq 0$ es té un esquema d'ordre de consistència 2 (com a mínim), que es denoten esquemes RK2.

Si $\alpha = 1$ s'obté el mètode

$$Y_{i+1} = Y_i + hf\left(Y_i + \frac{h}{2}f(Y_i,t_i), t_i + \frac{h}{2}\right)$$

que el que fa és utilitzar una aproximació de la derivada de y en el punt $t_i + h/2$ suposant que a t_i el valor de y és Y_i.

Si $\alpha = 1/2$ s'obté el mètode

$$Y_{i+1} = Y_i + h\frac{1}{2}\left(f(Y_i,t_i) + f\left(Y_i + hf(Y_i,t_i), t_i + h\right)\right)$$

que el que fa és fer la mitjana entre el pendent de y en t_i i una aproximació del pendent de y en $t_i + h$ suposant que a t_i el valor de y és Y_i.

Observeu que el paràmetre α és lliure, i té sentit preguntar-se si existeix algun valor d'α pel qual l'esquema sigui d'ordre 3. Si es té en compte el terme d'ordre 3 en la sèrie de potències (3.3) es pot veure que no hi ha cap valor α que cancel·li aquest terme (obviament utilitzant que c_1, c_2, a_2 i b_{21} estan donat en funció d'α per tal que l'esquema sigui almenys d'ordre 2). Per tant, per tot $\alpha \neq 0$ l'esquema que s'obté és d'ordre 2.

Esquemes Runge-Kutta generals

La idea anterior es pot generalitzar per construir esquemes de més etapes. Concretement, un esquema RK de m etapes és de la forma

$$Y_{i+1} = Y_i + h\phi(t_i, Y_i, h; f) \qquad \text{amb} \qquad \phi(t_i, Y_i, h; f) = \sum_{r=1}^{m} c_r K_r(t_i, Y_i, h; f)$$

on les funcions K_r estan donades per

$$K_r = f(Y_i + h\sum_{s=1}^{m} b_{rs}K_s, t_i + ha_r).$$

Els paràmetres c_r, a_r i b_{rs} per $r, s \in \{1, 2, \ldots, m\}$ se seleccionen de manera que l'esquema sigui d'ordre el més gran possible.

Observeu que si $b_{rs} = 0$ si $s \geq r$ aleshores el valor de $K_1, K_2, \ldots, K_m$ es pot calcular de forma explícita començant per K_1 i per ordre, concretament fent

$$\begin{aligned}
K_1 =& f(Y_i, t_i + ha_1) \\
K_2 =& f(Y_i + hb_{21}K_1, t_i + ha_2) \\
K_3 =& f(Y_i + h(b_{31}K_1 + b_{32}K_2), t_i + ha_3) \\
&\vdots \\
K_m =& f(Y_i + h(b_{m1}K_1 + hb_{m2}K_2 + \cdots + b_{m,m-1}K_{m-1}), t_i + ha_m)
\end{aligned}$$

Si $b_{rs} \neq 0$ per algun $s \geq r$, aleshores els valors de $K_1, K_2, \ldots, K_m$ no es poden calcular de forma explícita i s'ha de resoldre un sistema d'equacions no lineals per trobar els seus valors. En general s'utilitzen sistemes RK del primer tipus, anomenats esquemes RK explícits (en aquest b_{rs} es fixa a zero si $s \geq r$, i s'ajusten la resta de paràmtres per obtenir esquemes d'ordre el més gran possible).

Un esquema RK4 explícit (el 4 indica que té ordre de consistència igual a 4) molt utilitzat és l'esquema de 4 etapes:

$$Y_{i+1} = Y_i + h\frac{1}{6}(K_1 + 2K_2 + 2K_3 + K_4)$$

amb

$$\begin{aligned}
K_1 =& f(Y_i, t_i) \\
K_2 =& f\left(Y_i + \frac{h}{2}K_1, t_i + \frac{h}{2}\right) \\
K_3 =& f\left(Y_i + \frac{h}{2}K_2, t_i + \frac{h}{2}\right) \\
K_4 =& f(Y_i + hK_3, t_i + h)
\end{aligned}$$

Teorema (Butcher, 1964). Donat un esquema RK de m etapes explícit es té que l'ordre de consistència (p) màxim que es pot assolir és

m	1	2	3	4	5	6	7	8	9	$10 \leq m$
p	1	2	3	4	4	5	6	6	7	$p \leq m-2$

3.5 Control de pas

Fins ara hem considerat que per calcular $y(t_f)$ es discretitza l'interval $[t_0, t_f]$ en n passos de longitud $h = (t_f - t_0)/n$. Això, però, no és molt eficient perquè ens força a agafar passos molt petits tota l'estona si hi ha alguna regió de l'interval $[t_0, t_f]$ que requereix aquests passos petits. Per evitar això es pot utilitzar la informació local que dóna $\tau(h,t)$ per escollir la longitud del pas més adeqüada en cada situació (la idea és que s'utilitzaran passos "llargs" si $\tau(h,t)$ és petit i passos "curts" si $\tau(h,t)$ és gran). Vegem un primer exemple d'aquesta técnica.

Exemple. Considerem el problema

$$\begin{cases} y'(t) = f(y(t), t) \\ y(t_0) = y_0 \end{cases},$$

Volem resoldre el problema de manera que l'error que es produeix en cada pas sigui menor a un determinada tolerència $\varepsilon > 0$, és a dir es vol que

$$|y(t_i + h_i) - Y_{i+1}| < \varepsilon$$

on h_i és la longitud del pas que s'utilitza a l'iterat $i+1$ i $t_i = t_0 = \sum_{j=0}^{i-1} h_j$.

Suposem que utilitzem l'esquema d'Euler explícit a cada pas, és a dir que

$$Y_{i+1} = Y_i + hf(Y_i, t_i).$$

L'error de truncament local de l'esquema, definit com

$$\tau(h,t) = \frac{y(t+h) - y(t) - hf(y(t),t)}{h},$$

s'expressa en en sèrie de potències de h com

$$\tau(h,t) = \frac{h}{2} y''(t) + O(h^2).$$

Per tant, demanar que

$$y(t_i + h_i) - Y_{i+1} < \varepsilon$$

és com demanar que

$$h_i \tau(h_i, t_i) < \varepsilon$$

sota l'assumpció que $y(t_i) = Y_i$ (la qual cosa no és certa, però sí és cert que $y(t_i) \approx Y_i$ si l'esquema funciona, i per tant imposar aquesta darrera igualtat té sentit al menys des d'un punt de vista pràctic).

Si s'obvien els termes $O(h^2)$ de $\tau(h_i, t_i)$, la qual cosa té sentit si h és prou petit, aleshores la desigualtat anterior se satisfà si

$$\frac{h_i^2}{2} y''(t_i) < \varepsilon.$$

Observem finalment que

$$y''(t_i) = \partial_1 f(y(t_i), t_i) f(y(t_i), t_i) + \partial_2 f(y(t_i), t_i) = \partial_1 f(Y_i, t_i) f(Y_i, t_i) + \partial_2 f(Y_i, t_i)$$

on assumim novament que $y(t_i) = Y_i$. Per tant, per tal que $|y(t_i + h_i) - Y_{i+1}| < \varepsilon$ n'hi ha prou en prendre h_i que satisfaci

$$h_i < \sqrt{\frac{2\varepsilon}{\partial_1 f(Y_i, t_i) f(Y_i, t_i) + \partial_2 f(Y_i, t_i)}}.$$

A la pràctica es prendria el pas més llarg possible sempre que no sigui més gran que una determinat pas màxim $\hat{h}$ fixat a priori, és a dir es prendria

$$h_i = \min \left\{ \hat{h}, \sqrt{\frac{2\varepsilon}{\partial_1 f(Y_i, t_i) f(Y_i, t_i) + \partial_2 f(Y_i, t_i)}} \right\}.$$

És clar que l'exemple anterior es pot generalitzar per esquemes arbitraris. Vegem-ho considerant un esquema de la forma

$$Y_{i+1} = Y_i + h\phi(t_i, Y_i, h; f)$$

que té ordre de consistència p i que l'error de truncament s'expressa en sèrie de potències com

$$\tau(h, t) = d(t)h^p + O(h^{p+1}).$$

Aquesta informació es pot utilitzar per deteminar el pas més gran possible h_i que cal utilitzar a l'iterat $i + 1$ de manera que $|y(t_i + h) - Y_{i+1}| < \varepsilon$ per un $\varepsilon > 0$ donat. Concretament, com que

$$|y(t_i + h_i) - Y_{i+1}| = |h_i \tau(h_i, t_i)|,$$

n'hi ha prou (obviant el terme $O(h^{p+1})$ de l'error de truncament i assumint que $y(t_i) = Y_i$) en prendre h_i tal que

$$d(t) h_i^{p+1} < \varepsilon$$

és a dir

$$h_i < \sqrt[p+1]{\frac{\varepsilon}{d(t)}},$$

i com a l'exemple s'agafaria el mínim entre la part dreta de la desigualtat i un pas màxim $\hat{h}$ fixat a priori.

Observem que per utilitzar la tècnica de control de pas de l'esquema anterior cal conèixer el terme dominant de $\tau(h, t)$ quan s'expressa com a sèrie de potències respecte de h, concretament el factor

$d(t)$. Quan s'empren esquemes de Taylor és relativament fàcil calcular aquest factor (es faria essencialment com a l'exemple). Quan s'utilitzen altres esquemes, per exemple els esquemes RK, aleshores calcular el factor $d(t)$ no és tan fàcil. D'altra banda, l'avantatge dels esquemes RK és que es poden implementar sense necessitat de conèixer les derivades de la funció f, i això és quelcom que es perdria si es volgués calcular $d(t)$ com s'ha fet en l'exemple anterior, ja que per expressar $\tau(h,t)$ com a sèrie de potències de h caldria derivar la funció f diverses vegades.

Observació. Quan s'utilitzen mètodes de control de pas per calcular $y(t_f)$ a partir de $y(t_0)$ és pràcticament segur que en algun moment passarà que $t_i + h_i > t_f$. Quan això passa cal prendre h com $h = t_f - t_i$ per tal que Y_{i+1} representi l'aproximació de $y(t_f)$ (és a dir, per tal de no passar-nos de llarg de la coordenada t_f que és sobre la qual es vol calcular la funció y).

Càlcul automàtic del terme $d(t)$

Per evitar l'inconvenient anerior i calcular $d(t)$ de forma aproximada sense necessitat d'expandir en sèrie de potències de h la funció $\tau(h,t)$ es pot seguir la següent estrategia, que requereix utilitzar dos esquemes numèrics de diferent ordre (per qüestions d'eficiencia computacional el millor és que tinguin ordres de consistència consecutius). Siguin, per tant, els mètodes

$$Y_{i+1} = Y_i + h\phi(t_i, Y_i, h; f)$$

i

$$Y_{i+1} = Y_i + h\tilde{\phi}(t_i, Y_i, h; f)$$

els dos mètodes, i suposem que l'ordre de consistència del primer és p i el del segon és $p+1$. Això implica que

$$h\tau(h,t_i) = y(t_i+h) - Y_{i+1} = d(t_i)h^{p+1} + O(h^{p+2}),$$
$$h\tilde{\tau}(h,t_i) = y(t_i+h) - \tilde{Y}_{i+1} = \tilde{d}(t_i)h^{p+2} + O(h^{p+3}),$$

on

$$Y_{i+1} = y(t_i) + h\phi(t_i, y(t_i), h; f) \qquad \text{i} \qquad \tilde{Y}_{i+1} = y(t_i) + h\tilde{\phi}(t_i, y(t_i), h; f).$$

Observem, per tant, que si restem les dues equacions es cancel·la el terme $y(t_i+h)$ i es té que

$$\tilde{Y}_{i+1} - Y_{i+1} = d(t_i)h^{p+1} + O(h^{p+2}).$$

En particular, obviant el terme $O(h^{p+2})$ es pot aproximar $d(t_i)$ com

$$d(t_i) = \frac{\tilde{Y}_{i+1} - Y_{i+1}}{h^{p+1}}.$$

En aquesta expressió hem de pensar el pas $h = h_i$ com quelcom que s'ha determinat en l'iterat anterior (és a dir a l'iterat que ha permès calcular Y_i a partir de Y_{i-1}), i el valor $d(t_i)$ que dona l'expressió s'utilitza com a aproximació de $d(t_{i+1})$, de manera que el pas h_{i+1} que s'utilitza a l'iterat $i+2$ (l'iterat que calcula Y_{i+2} a partir de l'iterat Y_{i+1} s'ha d'agafar com

$$h_{i+1} = \min\left\{\hat{h},\ \sqrt[p+1]{\frac{\varepsilon}{|d(t_i)|}}\right\} = \min\left\{\hat{h}, h_i\sqrt[p+1]{\frac{\varepsilon}{|\tilde{Y}_{i+1} - Y_{i+1}|}}\right\},$$

on novament $\hat{h}$ seria un pas màxim fixat a priori. Observeu que per utilitzar aquesta técnica automàtica de control de pas cal donar un pas h_0 inicial (la resta de passos h_i per $i \geq 1$ es calculen utilitzant la fórmula anterior).

La tècnica anterior pot representar un cost computacional notable ja que s'estan utilitzant dos esquemes a cada pas. Existeixen parelles de mètodes RK, però, que tenen la propietat que per calcular $\tilde{Y}_{i+1}$ a partir de Y_i s'aprofiten tots els càlculs que es fan per calcular Y_{i+1} a partir de Y_i i només cal calcular una mica més per obtenir $\tilde{Y}_{i+1}$. La idea és que molts paràmetres del primer esquema s'aprofiten pel segon. Aquests esquemes combinats es coneixen com esquemes **Runge-Kutta-Fehlberg**. Vegem l'esquema RK2,3, format per un esquema RK2 (de tres etapes) i un altre RK3 (de 4 etapes) que s'obtenen com

$$\begin{aligned} Y_{i+1} &= Y_i + h\phi(t_i, Y_i, h; f) \quad \text{amb} \quad \phi(t_i, Y_i, h; f) = c_1K_1 + c_2K_2 + c_3K_3 \\ \tilde{Y}_{i+1} &= Y_i + h\tilde{\phi}(t_i, Y_i, h; f) \quad \text{amb} \quad \phi(t_i, Y_i, h; f) = \tilde{c}_1K_1 + \tilde{c}_2K_2 + \tilde{c}_3K_3 + \tilde{c}_4K_4 \end{aligned}$$

amb

$$\begin{aligned} K_1 &= f(Y_i, t_i) \\ K_2 &= f(Y_i + hb_{21}K_1, t_i + ha_2) \\ K_3 &= f(Y_i + h(b_{31}K_1 + b_{32}K_2), t_i + ha_3) \\ K_4 &= f(Y_i + h(b_{41}K_1 + b_{42}K_2 + b_{43}K_3), t_i + ha_4) \end{aligned}$$

Els paràmetres queden recollits a la següent taula:

r	a_r	c_r	$\tilde{c}_r$	b_{rs}		
1	0	214/891	533/2106	0	0	0
2	1/4	1/33	0	1/4	0	0
3	27/40	650/891	800/1053	−189/800	729/800	0
4	1	0	−1/78	214/891	1/33	650/891

Altres esquemes Runge-Kutta-Fehlberg força utilitzats són l'RK4,5 i l'RK7,8.

3.6 Regió d'estabilitat absoluta

Donat un esquema de la forma

$$Y_{i+1} = Y_i + h\phi(t_i, Y_i, h; f)$$

i un nombre $\lambda \in \mathbb{C}$, es diu que l'esquema associat al pas h és **abolutament estable** si l'esquema aplicat al PVI lineal

$$\begin{cases} y'(t) = \lambda y(t) \\ y(0) = y_0 \end{cases}$$

amb el pas h satisfà que $Y_i \to 0$ quan $i \to \infty$. Observeu que si $Re(\lambda) < 0$, aleshores la solució $y(t)$ del PVI va a zero quan t va a infinit, de manera que és desitjable que l'esquema numèric també exhibeixi aquesta propietat (és a dir, que si $Re(\lambda) < 0$ la solució numèrica Y_i tendeixi a zero quan i tendeix a infinit). En general això serà quelcom que passarà en mètodes d'un pas consistents si h és suficientment petit (ja que els mètodes monopas consistents són convergents, la qual cosa

pot no ser certa en mètodes mulipas consistents). Ens interessa, però, saber a partir de quin h l'esquema serà absolutament estable.

Exemple. Considerem el mètode d'Euler explícit, i apliquem aquest estquema al PVI anterior. La recurrència per la solució numèrica és

$$Y_{i+1} = Y_i + \lambda h Y_i.$$

És clar el terme general de la recurrència és

$$Y_i = (1 + h\lambda)Y_0,$$

de manera que $Y_i \to 0$ si i només si $|1 + h\lambda| < 1$. Si λ és real i negatiu, aquesta darrera inequació se satisfà si i només si

$$h < \frac{2}{|\lambda|}.$$

Per tant, si $\lambda = -200$ per exemple, la qual cosa implica que la solució $y(t)$ del PVI tendeix a zero molt ràpidament, es té que l'esquema d'Euler no convergirà a zero si $h \geq 2/200 = 0.01$ (de fet l'esquema divergirà a $\pm\infty$ si $h > 0.01$). Per tal que l'esquema exhibeixi el comportament adequat en aquest cas caldria agafar un pas h molt petit.

En molts esquemes (en particular els esquemes de Taylor i els RK) la recurrència associada a l'esquema quan s'aplica al PVI anterior és de la forma

$$Y_{i+1} = p(h\lambda)Y_i$$

on $p(z)$ és un polinomi (en l'exemple anterior $p(z) = 1 + z$). En aquests casos es té, per tant, que Y_i tendeix a zero si i només si $|p(h\lambda)| < 1$. Això motiva la definició de **regió d'estabilitat absoluta** d'aquests esquemes com el conjunt de $\mathbb{C}$

$$R_p := \{z \in \mathbb{C} \mid |p(z)| < 1\}.$$

Observeu que aquesta regió només depen de l'esquema considerat, i no dels valors λ i h utilitzats. Observeu també que la condició $|p(h\lambda)| < 1$ és certa si i només si $h\lambda \in R_p$, i per tant es té que l'esquema associat al pas h és absolutament estable pel PVI si $h\lambda \in R_p$. En particular, fixat λ el pas h més gran que es pot agafar seria el pas més gran pel qual $h\lambda \in R_p$.

Exercici. Considerem el PVI bidimensional

$$\begin{cases} x'(t) = -10x(t) - y(t) \\ y'(t) = x(t) - 10y(t) \\ x(0) = 1 \\ y(0) = 0 \end{cases}.$$

a) Cap a on tendeix la solució $t \mapsto (x(t), y(t))$ del PVI anterior quan t va a infinit?

b) Com de petit ha de ser h perquè la solució numèrica (X_i, Y_i) de l'esquema d'Euler explícit tendeixi a zero quan i tendeix a infinit?

Resolució. a) Com que es tracta d'un sistema d'EDOs lineal, la dinàmica està descrita pels valors propis de la matriu

$$A = \begin{pmatrix} -10 & -1 \\ 1 & -10 \end{pmatrix},$$

que són $\lambda_1 = -10 + i$ i $\lambda_2 = -10 - i$. Com que tots els valors propis tenen part real negativa, la solució del PVI tendeix al punt $(0,0)$ quan $t \to \infty$.

b) L'esquema d'euler explícit aplicat al problema anterior és de la forma

$$\begin{pmatrix} X_{i+1} \\ Y_{i+1} \end{pmatrix} = \begin{pmatrix} X_i \\ Y_i \end{pmatrix} + hA \begin{pmatrix} X_i \\ Y_i \end{pmatrix} = (\mathrm{Id} + hA) \begin{pmatrix} X_i \\ Y_i \end{pmatrix} = \begin{pmatrix} 1 - 10h & -h \\ h & 1 - 10h \end{pmatrix} \begin{pmatrix} X_i \\ Y_i \end{pmatrix}.$$

És clar que el terme general de successió associada a la solució numèrica que s'obté en aplicar l'iteració anterior és (en termes de la condició inicial)

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} = (\mathrm{Id} + hA)^n \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix}.$$

Per tal que el terme general anterior convergeix a zero quan el nombre d'iterats va a infinit (sigui quina sigui la condició inicial), cal que els valors propis de la matriu $\mathrm{Id} + hA$ tinguin mòdul més petit que 1. Com que els vectors propis de $(\mathrm{Id} + hA)$ són els mateixos que els vectors propis de A i els valors propis respectius són $\tilde{\lambda}_1 = 1 + h\lambda_1$ i $\tilde{\lambda}_2 = 1 + h\lambda_2$ (comproveu-ho), la solució tendirà a 0 (independentment de la condició inicial) només si $|1 + h\lambda_1| < 1$ i $|1 + h\lambda_2| < 1$, és a dir si $h\lambda_1$ i $h\lambda_2$ estan tot dos dins de la regió d'estabilitat associada al mètode d'Euler explícit.

Així, com que $|1 + h\lambda_1| = |1 - 10h + ih| = \sqrt{(1-10h)^2 + h^2}$ i $|1 + h\lambda_2| = |1 - 10h - ih| = \sqrt{(1-10h)^2 + h^2}$, cal imposar que $\sqrt{(1-10h)^2 + h^2} < 1$, que equival a demanar

$$(1 - 10h)^2 + h^2 < 1$$

i això és cert si $h < 20/101$.

Exercici. Considerem el PVI bidimensional

$$\begin{cases} x'(t) = a_{11}x(t) + a_{12}y(t) \\ y'(t) = a_{21}x(t) + a_{22}y(t) \\ x(0) = 1 \\ y(0) = 0 \end{cases},$$

on

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} -50 & 0 \\ 0 & -0.1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}^{-1}.$$

a) Cap a on tendeix la solució $t \mapsto (x(t), y(t))$ del PVI anterior quan t va a infinit?

b) Com de petit ha de ser h perquè la solució numèrica (X_i, Y_i) de l'esquema d'Euler explícit tendeixi a zero quan i tendeix a infinit?

c) Com de petit ha de ser h perquè la solució numèrica (X_i, Y_i) de l'esquema d'Euler implícit tendeixi a zero quan i tendeix a infinit?

Exercici. Considereu la família d'esquemes

$$\begin{cases} Y_{n+1} = Y_n + h(\alpha f_n + (1-\alpha) f_{n+1}) \\ X_0 = x_0 \end{cases},$$

on $f_n := f(Y_n, t_n)$ amb $t_n = nh$ i on $\alpha \in [0,1]$, associats al problema

$$\begin{cases} y'(t) = f(y(t), t) \\ y(0) = y_0 \end{cases}.$$

a) Proveu que en general els esquemes són d'ordre 1. Trobeu un valor α que faci que l'esquema tingui ordre 2.

b) Per aquest α particular, preneu $f(y,t) = \lambda y$, amb $\lambda \in \mathbb{C}$. Proveu que l'esquema, independentment del pas de temps h, divergeix si $\mathrm{Re}(\lambda) > 0$ i convergeix a 0 si $\mathrm{Re}(\lambda) < 0$ (observeu que aquesta propietat no la satisfà ni el mètode d'Euler explícit ni el mètode d'Euler implícit, que es corresponen als casos $\alpha = 1$ i $\alpha = 0$ respectivament).

Exercici. Considereu el següent mètode multipas per a l'equació $y'(t) = f(y(t), t)$:

$$Y_{n+1} = Y_n + \frac{h}{2}(3f_n - f_{n-1})$$

on $f_n := f(Y_n, t_n)$. Proveu que l'error de truncament local és d'ordre 2, i.e. que $\tau = O(h^2)$.

Resolució. Es procedeix igual que pels mètodes d'un sol pas. L'error de truncament es defineix com la diferència entre la solució real i la solució que dóna el mètode (quan s'aplica al punt t), i tot això dividit pel pas h, és a dir

$$\tau(h,t) = \frac{1}{h}(y(t+h) - y(t) - \frac{h}{2}(3f(y(t),t) - f(y(t-h), t-h)).$$

Com que $f(y(t),t) = y'(t)$ i $f(y(t-h), t-h) = y'(t-h)$, per escriure $\tau(h,t)$ en sèrie de potències (fins a ordre 2, que és el que ens demanen), utilitzem que

$$y(t+h) = y(t) + y'(t)h + y''(t)\frac{h^2}{2} + y'''(t)\frac{h^3}{6} + O(h^4),$$
$$y'(t-h) = y'(t) - y''(t)h + y'''(t)\frac{h^2}{2} + O(h^3),$$

de manera que

$$\tau(h,t) = y'''(t)h^2\left(\frac{1}{6} + \frac{1}{4}\right) + O(h^3),$$

i per tant el mètode és d'ordre 2 (ja que el terme dominant no s'anul·la en general).

Exercici. Recordeu que els mètodes de Runge-Kutta d'm etapes són de la forma:

$$Y_{n+1} = Y_n + h\sum_{i=1}^{m} b_i K_i$$

amb

$$K_i = f\left(Y_n + h\sum_{j=1}^{m} a_{ij}K_j, t_n + c_i h\right).$$

Proveu que aquests mètodes són consistents si i només si

$$\sum_{i=1}^{m} b_i = 1$$

Resolució. Es fa de forma similar als altres exercicis en que s'ha calculat l'error de truncament (per aquest exercici n'hi ha prou en escriure la sèrie de potències de $\tau(h,t)$ fins a ordre 0, i imposar que el terme d'ordre 0 és zero, la qual cosa implica que el mètode és almenys d'ordre 1 i per tant consistent).

Exercici. Recordem la famíla d'esquemes RK2 (i.e. d'ordre 2) de la forma

$$Y_{i+1} = Y_i + h(c_1K_1(Y_i,t_i) + c_2K_2(Y_i,t_i))$$

amb

$$K_1(Y_i,t_i) = f(Y_i,t_i)$$
$$K_2(Y_i,t_i) = f(Y_i + hb_{21}K_1(Y_i,t_i), t_i + ha_2)$$

on

$$c_2 = \alpha, \quad c_1 = 1-\alpha, \quad a_2 = \frac{1}{2\alpha} \quad \text{i} \quad b_{21} = \frac{1}{2\alpha}.$$

Estudieu si la regió d'estabilitat absoluta dels esquemes anteriors depèn del paràmetre α o no.

Resolució. Cal analitzar com és la recurrència de la solució numèrica quan s'aplica el mètode a un porblema de la forma

$$\begin{cases} y'(t) = \lambda y(t) \\ y(0) = y_0 \end{cases}.$$

Observem que per aquest problema es té que $f(y,t) = \lambda y$, i per tant

$$K_1(Y_i,t_i) = \lambda Y_i$$
$$K_2(Y_i,t_i) = \lambda(Y_i + hb_{21}\lambda Y_i) = \lambda(1+hb_{21}\lambda)Y_i.$$

Utilitzant ara l'expressió del mètode i el valor dels coeficients a, c i b es té

$$Y_{i+1} = Y_i + (1-\alpha)h\lambda Y_i + \alpha h\lambda\left(1+\frac{1}{2\alpha}h\lambda\right)Y_i = \left(1+(1-\alpha)h\lambda + \alpha h\lambda\left(1+\frac{1}{2\alpha}h\lambda\right)\right)Y_i$$

Per tant, definint

$$p(z) = \left(1+(1-\alpha)z + \alpha z\left(1+\frac{1}{2\alpha}z\right)\right)$$

es té que la regió d'estabilitat absoluta del mètode és el conjunt

$$R_p := \{z \in \mathbb{C} \mid |p(z)| < 1\}.$$

Ara bé, desenvolupant el polinomi p s'observa que p no depen del paràmetre α:

$$p(z) = \left(1+z+\frac{1}{2}z^2\right),$$

i per tant es conclou que la regió d'estabilitat absoluta de la família de mètodes anteriors no depen del paràmetre α.