Pràctica 3. Mètodes de Monte Carlo

1 Introducció

En aquesta pràctica veurem dos mètodes de Monte Carlo per a calcular integrals (definides) mitjançant nombres aleatoris. Per simplificar la pràctica, ens limitarem al cas unidimensional, tot i que en aquest cas, en general, es poden trobar mètodes numèrics deterministes molt millors (més ràpids, amb millor control de l'error) que els estocàstics. Els mètodes de Monte Carlo són realment útils per a calcular integrals multidimensionals, on els mètodes deterministes poden ser inaplicables; però en aquest cas el mètode de Monte Carlo groller (*Crude Monte Carlo*) que nosaltres estudiarem s'ha de refinar per tal que sigui eficient, i aquests refinaments queden fora de l'abast d'aquest curs.

2 Càlcul d'una integral pel mètode d'encertar o fallar (*hit-or-miss)*

Volem calcular una integral definida

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

on $-\infty < a < b < \infty$, i la funció g és contínua i positiva; g estarà afitada en la regió d'integració: per algun $c > 0$ (com més ajustada sigui c millor anirà el mètode):

$$0 \le g(x) \le c, \quad \forall x \in [a, b].$$

(Si g no és positiva, però g^+ i g^- tenen bones propietats, el mètode es pot aplicar a cada part per separat). Anomenen A a la regió sota la corba $y = g(x)$ entre els punts a i b. La integral I és igual a l'àrea de la regió A. Vegeu la Figura 1.

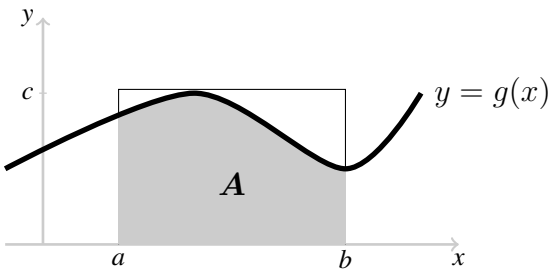


Figura 1. La integral I és igual a l'àrea de la regió A

Considerem un vector aleatori (X, Y) uniforme en $(a, b) \times (0, c)$, i sigui $f(x, y)$ la seva funció de densitat conjunta:

$$f(x, y) = \frac{1}{c(b-a)} \mathbf{1}_{(a,b)\times(0,c)}(x, y).$$

Noteu que les variables X i Y són independents. Tenim que

$$p := P\{(X,Y) \in A\} = \iint_A f(x,y)\,dxdy = \frac{1}{c(b-a)} \iint_A dxdy = \frac{\text{àrea d'}A}{c(b-a)} = \frac{I}{c(b-a)}. \qquad (1)$$

Així,

$$I = c(b-a)\,p.$$

Per calcular aproximadament aquesta integral prenem punts $(X_1,Y_1),\ldots,(X_n,Y_n)$ independents aleatòriament sobre el rectangle $(a,b)\times(0,c)$ i comptem quants estan a la zona grisa A. Noteu que

$$(X_j,Y_j) \in A \quad \Longleftrightarrow \quad Y_j \le g(X_j).$$

Designem per p_n la freqüència relativa d'èxits:

$$p_n = \frac{\#\{j:\ Y_j \le g(X_j)\}}{n} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{(X_j,Y_j)\in A\}}, \qquad (2)$$

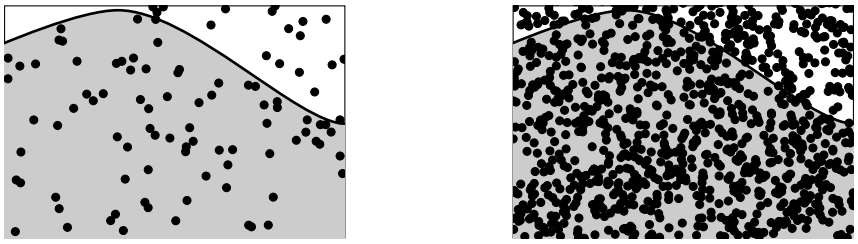
on $\#B$ designa el cardinal d'un conjunt B. Per la llei dels grans nombres, per a n gran,

$$p_n \approx p,$$

i per tant,

$$I_n := c(b-a)p_n \approx c(b-a)p = I.$$

Vegeu la Figura 2.



(a) $n = 100$ (b) $n = 1000$

Figura 2. El mètode d'encertar o fallar

Exemple 1.

Càlcul per Monte Carlo de

$$\int_0^1 \exp(e^x)\,dx$$

amb $n = 10^4$. Càlcul numèric amb l'**R** d'aquesta integral i comparació dels dos valors.

Abans de començar hem de trobar el valor c tal que

$$\exp(e^x) \le c,\ \forall x \in [0,1].$$

En aquest cas la funció és creixent i el màxim de la funció es troba en el punt $x = 1$, i $c = e^e \approx 15.15$.

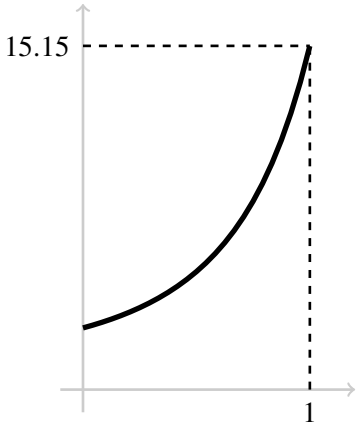


Figura 3. Funció $g(x) = e^{e^x}$.

```
> library(MASS)                 # necessitem carregar aquest paquet
                                # per a fer la integració numèrica
> g=function(x){exp(exp(x))}
> n=10^4
> c=exp(exp(1))                 # fita
> x=runif(n)
> y=c*runif(n)                  # generem n uniformes a  (0,c)
                                # degut a que X i Y són independents
                                # podem generar-les per separat
                                # i considerar el vector (X,Y) amb
                                # la llei que volem
> Imonte=c*sum(y<g(x))/n        # estimació de Monte Carlo
> Inum=area(g,0,1)              # càlcul numeric
> error=abs(Inum-Imonte)
```

Si la fita c per a la funció g a l'interval $[a, b]$ s'ha de calcular numèricament es pot utilitzar la instrucció `c=optimize(g,interval=c(a,b),maximum=T)$objective`, on prèviament cal haver definit la funció g.
Aquí `...$objective` vol dir que dels resultats que proporciona la funció `optimize` seleccionem el màxim de la funció.

Problema 1

Calculeu per Monte Carlo ($n = 10^4$) la integral

$$\int_0^1 (1 - x^2)^{3/2}\, dx.$$

Calculeu amb l'**R** un valor numèric d'aquesta integral i compareu ambdós resultats.

2.1 Avaluació de l'error

Amb les notacions que hem introduït abans (vegeu (1) i (2)),

$$\mathbf{E}[I_n] = c(b-a)\mathbf{E}[p_n] = c(b-a)\mathbf{E}\big[\mathbf{1}_{\{(X,Y)\in A\}}\big] = c(b-a)\mathbf{P}\big\{(X,Y)\in A\big\} = c(b-a)p = I,$$

la qual cosa vol dir que si fem moltes vegades el càlcul d'I_n, en mitjana el resultat serà correcte. Es diu que I_n és un estimador **sense biaix** d'I. D'altra banda, d'acord amb el Teorema del Límit Central, l'error $|I - I_n|$ estarà controlat per la desviació típica de I_n. Calculem la variància de I_n

$$\mathrm{Var}(I_n) = \frac{1}{n}\, c^2(b-a)^2\big(p - p^2\big) = \frac{1}{n}\big[c(b-a)I - I^2\big]. \tag{3}$$

De fet, estem en la situació que havíem estudiat a la pràctica 2, i l'error és del tipus $O(n^{-1/2})$.

3 Càlcul d'integrals mitjançant una esperança: el mètode de mostrejar allí on és important *(importance sampling)*

En aquesta pràctica utilitzarem la següent propietat:

Sigui X una variable aleatòria amb funció de densitat $f(x)$ i h una funció tal que $h(X)$ tingui esperança finita. Aleshores,

$$\boxed{\mathbf{E}[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)\,dx.} \tag{4}$$

També utilitzarem la següent conseqüència de la llei forta dels grans nombres:

Sigui Z una variable aleatòria amb esperança finita, i $Z_1, Z_2, \ldots$ una successió de variables i.i.d. totes amb la llei de Z. Aleshores,

$$\lim_n \frac{1}{n}\sum_{j=1}^{n} Z_j = \mathbf{E}[Z],\ \text{q.s.}$$

En particular, si ho apliquem a la situació anterior amb $h(x)$ tindrem que si $X_1, \ldots, X_n$ són i.i.d. amb llei X, llavors

$$\boxed{\frac{1}{n}\sum_{j=1}^{n} h(X_j) \approx \mathbf{E}[h(X)].} \tag{5}$$

Normalment calculem l'esperança d'una funció d'una variable aleatòria contínua mitjançant una integral ordinària, d'acord amb (4). En aquest mètode de Monte Carlo utilitzarem el camí invers i escriurem una integral com l'esperança d'una funció d'una variable aleatòria, la qual aproximarem per una mitjana mostral com a (5). Com abans, l'objectiu és calcular

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

però ara a o b poden ser infinits, $-\infty \leq b < a \leq \infty$, i no cal demanar $g(x) \geq 0$. La idea és expressar la integral de la següent manera:

$$I = \int_a^b \frac{g(x)}{f(x)}\, f(x)\, dx,$$

on f és una funció de densitat en (a, b), i cal que $f(x) > 0$ en (a, b). Escrivim

$$h(x) = \frac{g(x)}{f(x)}.$$

Aleshores

$$I = \int_a^b g(x)\, dx = \int_a^b h(x) f(x)\, dx = \mathsf{E}[h(X)],$$

on X té densitat f. La densitat f s'anomena **densitat instrumental** (també s'utilitzen altres noms com *densitat candidata*). Ara considerem una mostra $X_1, \ldots, X_n$ d'una llei amb densitat f i definim

$$I_n^e = \frac{1}{n} \sum_{j=1}^n h(X_j).$$

(Utilitzem la notació I_n^e per distingir-lo del mètode anterior.) Tal com hem dit, degut a la llei forta dels grans nombres

$$\lim_n I_n^e = I.$$

Noteu també que

$$\mathsf{E}[I_n^e] = \frac{1}{n} \sum_{j=1}^n \mathsf{E}[h(X_j)] = \mathsf{E}[h(X)] = I. \tag{6}$$

És a dir, I_n^e és un estimador sense biaix d'I.

* * *

Exemple. Volem calcular

$$I = \int_5^\infty x^2\, e^{-x}\, dx = \int_0^\infty x^2\, e^{-x}\, \mathbf{1}_{(5,\infty)}(x) dx.$$

Aquí la funció g serà

$$g(x) = x^2\, e^{-x}\, \mathbf{1}_{(5,\infty)}(x).$$

Sembla raonable utilitzar com a densitat instrumental la densitat d'una exponencial de paràmetre 1: $f(x) = e^{-x} \mathbf{1}_{(0,\infty)}(x)$. Llavors prenem $h(x) = x^2 \mathbf{1}_{(5,\infty)}(x)$ i

$$I = \int_0^\infty x^2 \mathbf{1}_{(5,\infty)}(x) e^{-x}\, dx = \mathsf{E}\big[X^2 \mathbf{1}_{(5,\infty)}(X)\big],$$

on $X \sim \mathcal{E}xp(1)$. Així, per a fer el càlcul, generarem n variables independents exponencials de paràmetre 1, $X_1, \ldots, X_n$, i calcularem

$$I_n^e = \frac{1}{n} \sum_{j=1}^n X_j^2 \mathbf{1}_{(5,\infty)}(X_j).$$

Exemple 2. Càlcul aproximat de la integral

$$I = \int_5^\infty x^2\, e^{-x}\, dx$$

simulant $n = 10^4$ lleis exponencials de paràmetre 1 amb `rexp(n)`. Calcularem també el valor de la integral que dóna l'**R** amb la instrucció `integrate` i calcularem l'error.

```
> h=function(x){x^2*(x>5)}
> n=10^4
> Imonte=sum(h(rexp(n)))/n
> g=function(x){x^2*exp(-x)}
> Inum=integrate(g,5,Inf)  # utilitzem aquesta instrucció en lloc d'àrea
> error=abs(Inum$value-Imonte)
```

Nota. Una exponencial de paràmetre 1 és pot obtenir com $-ln(U)$, on U és una variable uniforme $\mathcal{U}(0,1)$. De manera que no caldria utilitzar la instrucció per generar exponencials, i en tindríem prou generant variables uniformes en $(0,1)$.

* * *

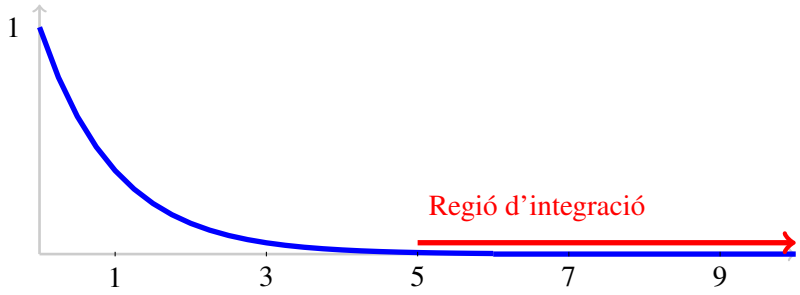


Figura 4. Densitat exponencial $f(x) = e^{-x}$, $x > 0$

Continuació de l'exemple. Donat que estem fent una integral sobre $(5,\infty)$ i la densitat de la exponencial dóna molta massa a l'interval (0,5), que es desaprofita (vegeu la Figura 4), seria molt millor utilitzar una densitat que estès concentrada en $(5,\infty)$, de la forma

$$f_1(x) = e^{-x+5}\mathbf{1}_{(5,\infty)}(x).$$

(vegeu la Figura 5)

Aquesta densitat correspon a una exponencial de paràmetre 1 però desplaçada a la dreta 5 unitats.

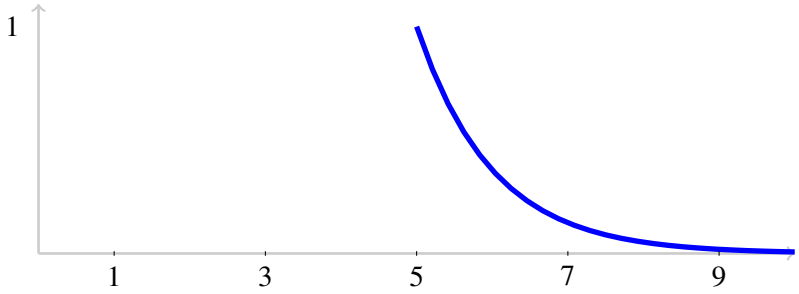


Figura 5. Densitat $f_1(x) = e^{-x+5},\ x > 5$.

Problema 2

Utilitzeu les fórmules del càlcul de la funció de densitat d'una variable aleatòria amb densitat per a demostrar que si $X \sim \mathcal{E}xp(1)$, aleshores

$$Y := X + 5$$

té la densitat f_1.

Llavors

$$I = \int_5^\infty x^2\, e^{-x}\, dx = e^{-5} \int_5^\infty x^2 f_1(x)\, dx = e^{-5}\mathbf{E}[Y^2] = e^{-5}\mathbf{E}[(X+5)^2],$$

on $X \sim \mathcal{E}xp(1)$. Per tant podem calcular la integral aproximada a partir d'una mostra $X_1, \ldots, X_n$ d'exponencials de paràmetre 1,

$$I_n^{e,1} := e^{-5}\frac{1}{n}\sum_{j=1}^{n}(X_j + 5)^2.$$

Problema 3

Calculeu $I_n^{e,1}$ amb $n = 10^4$. Noteu que, tot i que la funció g no ha canviat, ara no cal utilitzar cap indicador. Utilitzant de nou el valor exacte que dóna l'**R** d'aquesta integral, calculeu l'error.

3.1 Escollint entre dues densitats instrumentals

Hem vist dues possibilitats per a calcular aproximadament la integral

$$I = \int_5^\infty x^2\, e^{-x}\, dx.$$

Els resultats de cada simulació, I_n^e i $I_n^{e,1}$, són variables aleatòries.

Problema 4

Calculeu 10 valors de la integral (cadascun amb $n = 10^4$)

$$I = \int_5^\infty x^2\, e^{-x}\, dx$$

amb la densitat f i 10 valors amb la densitat f_1. Noteu que els primers 10 valors estan més dispersos que els 10 obtinguts amb la segona densitat. Per exemple, podeu representar en una recta els valors calculats amb les dues densitats (uns en color vermell, els altres en color blau). Representeu en aquesta recta el valor exacte (en negre).

D'acord amb (6)

$$\mathsf{E}[I_n^e] = \mathsf{E}[I_n^{e,1}] = I,$$

és a dir, ambdós I_n^e i $I_n^{e,1}$ són estimadors sense biaix d'I. Però segons hem vist al problema anterior, els valors de I_n^e estan més dispersos que els de $I_n^{e,1}$. Atès que normalment només es fa el càlcul un cop, l'estimador que tingui menys variabilitat serà el que anirà millor. Més concretament, tenim

$$\mathrm{Var}(I_n^e) = \frac{1}{n}\mathrm{Var}(h(X)) = \frac{1}{n}\Big(\mathsf{E}\big[\big(h(X)\big)^2\big] - \big(\mathsf{E}[h(X)]\big)^2\Big) = \frac{1}{n}\Big(\int_a^b \frac{g^2(x)}{f(x)}\, dx - I^2\Big). \qquad (7)$$

Per tant veiem que aquest mètode també és $O(n^{-1/2})$. Definim

$$\sigma_f^2 = \int_a^b \frac{g^2(x)}{f(x)}\, dx - I^2. \qquad (8)$$

Aleshores la desviació típica de I_n^e és $\sigma_f/\sqrt{n}$. Per tant, la qualitat de l'aproximació dependrà de σ_f. Utilitzant que per a una variable aleatòria Y sempre es té

$$\mathsf{E}[Y^2] \geq \big(\mathsf{E}[Y]\big)^2$$

pot argumentar-se que la millor densitat f serà aquella que sigui proporcional a $|g(x)|$. A la pràctica aquesta densitat acostuma a ser força difícil de trobar. Buscar una bona densitat que minimitzi σ_f^2 és un dels mètodes anomenats de **reducció de la variància**.

Retornant a l'exemple 2 i al problema 3, podeu calcular amb l'**R** (fent les integrals numèricament) les quantitats σ_f^2 i $\sigma_{f_1}^2$ i comprovar que $\sigma_{f_1}^2 < \sigma_f^2$. Llavors, el càlcul amb f_1 dóna resultats més concentrats al voltant de la mitjana que el càlcul amb f, que és el que havíem observat.

3.2 Comparació amb el mètode d'encertar o fallar

La integral que hem calculat al problema 1

$$\int_0^1 \exp(e^x)\, dx$$

també podem calcular-la per *importance sampling* utilitzant com a densitat instrumental, per exemple, la densitat d'una uniforme $U \sim \mathcal{U}(0,1)$:

$$\int_0^1 \exp(e^x)\,dx = E[\exp(e^U)] \approx \frac{1}{n}\sum_{j=1}^{n} \exp(e^{U_j}),$$

on $U_1, \ldots, U_n$ són i.i.d. $\mathcal{U}(0,1)$. En comparació amb el mètode d'encertar o fallar, ara només hem de simular una llei uniforme a (0,1) n vegades, amb la qual cosa ja hi ha un enorme guany de temps.

Per analitzar quin mètode és millor, escrivim en general la situació que hem descrit a la secció 1: tenim a, b finits, i $0 \leq g(x) \leq c$ i volem calcular

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

Anomenen I_n la integral aproximada que obtenim utilitzant el mètode d'encertar o fallar. Utilitzem importance sampling amb una densitat instrumental $\mathcal{U}(a,b)$, i anonemen I_n^e a la integral aproximada obtinguda; d'acord amb la fórmula (7),

$$\text{Var}(I_n^e) = \frac{1}{n}\Big(\int_a^b \frac{g^2(x)}{f(x)}\,dx - I^2\Big) = \frac{1}{n}\Big((b-a)\int_a^b g^2(x) - I^2\Big). \tag{9}$$

Ara comparem (9) amb la variància que havíem obtingut amb el mètode d'encertar o fallar (3), i veiem que el punt està en comparar

$$\int_a^b g^2(x)\,dx \qquad \text{i} \qquad c\int_a^b g(x)\,dx.$$

Però donat que $0 \leq g(x) \leq c,\ \forall c \in (a,b)$, tindrem

$$\int_a^b g^2(x)\,dx \leq c\int_a^b g(x)\,dx.$$

Per tant

$$\text{Var}(I_n^e) \leq \text{Var}(I_n).$$

Per tant, a la pràctica mai s'ha d'utilitzar el métode d'encertar o fallar.

Problema 5

Retornant al problema 1, calculeu la integral $\int_0^1 \exp(e^x)\,dx$ utilitzant el mètode d'*importance sampling* amb una densitat instrumental uniforme a $(0,1)$, i calculeu l'error.