Véletlenszám-generátorok

5. rész

– egy MATLAB® alapú megközelítés –

Baja Zsolt, Vas Orsolya

Matematika és Informatika Intézet, Babeș–Bolyai Tudományegyetem, Kolozsvár, Románia

(bajazsolt98@gmail.com, vas.orsolya@yahoo.com)

5. labor / 2023. október 30. – november 3.

- Az elutasítás módszerének alapértelmezett változata azt feltételezte, hogy az $f : \mathbb{R}^d \to [0,1]$ sűrűségfüggvényű $X : \Omega \to \mathbb{R}^d$ valószínűségi vektor esetén létezik egy olyan $g : \mathbb{R}^d \to [0,1]$ sűrűségfüggvény és egy olyan $c \geq 1$ elutasítási konstans, amelyek esetén teljesül az
$$f(x) \leq cg(x), \forall x \in \mathbb{R}^d \tag{1}$$
egyenlőtlenség.
- A 4. labor számgenerátoraiban, az (1)-es egyenlőtlenségre épülő elutasítás módszerének alkalmazása során, az $\frac{f}{cg}$ arányt N-szer értékeltük ki, ahol az N geometriai eloszlású valószínűségi változó a bemutatott algoritmusok **ismételd–ameddig** ciklusainak lépésszámát mérte. Azt is kimutattuk, hogy az N várható értékét éppen az elutasítási konstans határozza meg, azaz
$$E(N) = c.$$
- Általában az $\frac{f}{cg}$ arány kiértékelése egy lassú és költséges művelet, mert az f sűrűségfüggvény alakja feltehetően bonyolult, ellenben találtunk volna más módszert is ilyen sűrűségfüggvényű valószínűségi változók mintavételezésére. Mi több előfordulhat, hogy az f explicit alakja nem ismert, mert vagy integrálalakban, vagy valamilyen nemlineáris egyenlet megoldásaként adott. Ilyen esetekben célszerű hatékonysági szempontok miatt az $\frac{f}{cg}$ tört kiértékelését vagy teljesen elkerülni, vagy használatát minimálisra csökkenteni.
- Ebben a részben a George Marsaglia által bevezetett [Marsaglia, 1977] közrefogás módszerét ismertetjük, amely elkerüli az $\frac{f}{cg}$ arány nagy valószínűséggel történő kiértékelését.

- A módszer valójában abból áll, hogy egy adott $f : \mathbb{R}^d \to [0,1]$ sűrűségfüggvény esetén találjunk két olyan könnyen kiértékelhető $h_1, h_2 : \mathbb{R}^d \to \mathbb{R}$ függvényt, amelyre fennáll a

$$0 \leq h_1(x) \leq f(x) \leq h_2(x) \leq cg(x),\ \forall x \in \mathbb{R}^d \tag{2}$$

egyenlőtlenség.

- Ha a (2)-es feltételt sikerül biztosítani, akkor az alábbi algoritmushoz hasonlóan járhatunk el f sűrűségfüggvényű valószínűségi változók generálására.

1. Algoritmus (A közrefogás módszere)

- *Bemenet:* a (2)-es feltételt kielégítő $f, g : \mathbb{R}^d \to [0,1]$ sűrűségfüggvények és a $h_1, h_2 : \mathbb{R}^d \to \mathbb{R}$ határoló függvények, valamint az $n \geq 1$ természetes szám, amely a kimeneti mintavétel méretét határozza meg.
- *Kimenet:* az f sűrűségfüggvényű valószínűségi változónak egy n-elemű független $\{X_i\}_{i=1}^n$ mintavétele.

1. **Minden** $i = 1, 2, \ldots, n$ index esetén **végezd el**:
2. **ismételd:**
3. generáld az $U \sim \mathcal{U}([0,1])$, valamint az ettől független g sűrűségfüggvényű Y valószínűségi változót
4. $W \leftarrow Ucg(Y)$
5. $L \leftarrow [W \leq h_1(Y)]$
6. **ha** $[L == \text{hamis}]$
7. **akkor ha** $W \leq h_2(Y)$, **akkor** $L \leftarrow [W \leq f(Y)]$
8. **ameddig** $[L == \text{igaz}]$
9. $X_i \leftarrow Y$.

- Az 1. algoritmusban egy L nevű logikai változót vezettünk be arra számítva, hogy igaz értéket legtöbbször a $W \leq h_1(Y)$ összehasonlítás (úgynevezett *gyors elfogadási lépés*) során vesz fel. A fennmaradt esetekben az úgynevezett *gyors elutasítási lépést* ($W > h_2(Y)$) használjuk, továbbá csak azon ritka esetekben értékeljük ki a költséges $W \leq f(Y)$ összehasonlítást, amikor $h_1(Y) \leq W \leq h_2(Y)$.
- A gyors elfogadási és gyors elutasítási lépések elhagyásával a klasszikus elutasítás módszeréhez jutunk vissza a $W = Ucg(Y) \leq f(Y)$ leállási feltétel alkalmazásával. Vegyük észre, hogy minden egyes minta generálása várhatóan $E(N) = c$ lépésig tart, akárcsak a klasszikus elutasítás módszerére alapuló algoritmusok esetén!
- A továbbiakban az N_f valószínűségi változót tanulmányozzuk, mely azt számolja, hogy hány alkalommal értékeltük ki a költséges f sűrűségfüggvényt egy minta generálásakor.

- Jelölje W_j egy adott minta generálásakor az **ismételd–ameddig** ciklus j-edik iterációjában ($j = 1, 2, \ldots, N$) kapott W valószínűségi változót, valamint legyen Y_j az ugyanebben az iterációban generált g sűrűségfüggvényű Y valószínűségi változó!
- Ekkor az

$$N_f = \sum_{j=1}^{N} \mathbb{1}_{[h_1(Y_j) < W_j < h_2(Y_j)]} \tag{3}$$

egyenlőséget kapjuk, ahol $\mathbb{1}_A : \Omega \to \{0, 1\}$ az A eseményhez tartozó indikátor függvényt jelöli, azaz

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \text{ha } \omega \in A, \\ 0, & \text{ha } \omega \notin A. \end{cases}$$

- A továbbiakban a következőképpen gondolkodhatunk:

$$\begin{aligned} E(N_f) &= E(N)\,E\left(\mathbb{1}_{[h_1(Y_1)<W_1<h_2(Y_1)]}\right) \\ &= E(N)\,P\left(h_1(Y_1) < W_1 < h_2(Y_1)\right) \\ &= E(N)\int_{\mathbb{R}^d} P\left(h_1(Y_1) < W_1 < h_2(Y_1) \mid Y_1 = y\right) g(y)\,\mathrm{d}y \\ &= c\int_{\mathbb{R}^d} P\left(h_1(y) < Ucg(y) < h_2(y)\right) g(y)\,\mathrm{d}y \\ &= c\int_{\mathbb{R}^d} \frac{h_2(y) - h_1(y)}{cg(y)} g(y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} \left(h_2(y) - h_1(y)\right) \mathrm{d}y, \end{aligned}$$

ahol az utolsó előtti lépésében felhasználtuk a (2)-es egyenlőtlenséget, valamint azt a tényt, hogy az U valószínűségi változó egyenletes eloszlású a $[0,1]$ intervallumon.

- Amennyiben $h_1(x) \equiv 0$ és $h_2(x) \equiv cg(x)$, vagyis az f sűrűségfüggvényt nem fogjuk közre, akkor
$$E(N_f) = E(N) = c,$$
mert a klasszikus elutasítás módszeréhez kerülünk vissza.
- Ha viszont csak egy gyors elfogadási lépésünk van, azaz ha
$$0 \leq h_1(x) \leq f(x) \leq h_2(x) = cg(x), \forall x \in \mathbb{R}^d,$$
akkor
$$E(N_f) = c - \int_{\mathbb{R}^d} h_1(x)\,\mathrm{d}x.$$

- Egy $f : \mathbb{R} \to [0,1]$ alakú sűrűségfüggvény közrefogásához szükséges könnyen kiértékelhető alsó és felső becslőfüggvények szerkesztéséhez például az f Taylor-féle sorbafejtését használhatjuk.
- Ha az f sűrűségfüggvény n-szer folytonosan deriválható, akkor

$$f(x) = f(0) + \frac{x}{1!} f^{(1)}(0) + \frac{x^2}{2!} f^{(2)}(0) + \ldots + \frac{x^{n-1}}{(n-1)!} f^{(n-1)}(0) + \frac{x^n}{n!} f^{(n)}(\xi), \quad (4)$$

ahol ξ az x paraméter előjelének függvényében egy $[0,x]$, vagy $[x,0]$ intervallumbeli szám, és $f^{(r)}(x) = \frac{\mathrm{d}^r}{\mathrm{d}x^r} f(x)$, $r = 0, 1, \ldots, n$.

- A (4)-es sorbafejtés utolsó tagját tanulmányozva, az f sűrűségfüggvény számára polinomiális alsó és felső becslőfüggvényeket szerkeszthetünk.

- Példaként térjünk vissza a standard normális eloszlás

$$f_{\mathcal{N}(0,1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},\ x \in \mathbb{R}$$

  sűrűségfüggvényéhez!

- Vegyük észre, hogy az

$$e^{-x} = 1 - x + \frac{x^2}{2!} - \ldots + (-1)^{n-1} \frac{x^{n-1}}{(n-1)!} + (-1)^n \frac{x^n}{n!} e^{-\xi}$$

  függvény minden $x \geq 0$ érték esetén a jól ismert

$$e^{-x} = \sum_{j=0}^{\infty} (-1)^j \frac{x^j}{j!}$$

  sorbafejtés két egymás utáni tagja között húzódik!

- Ezért sajátosan teljesül az

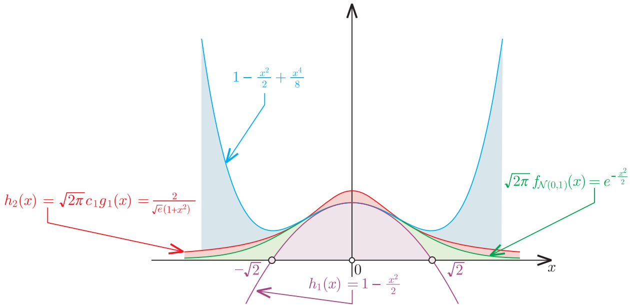
$$1 - x \leq e^{-x} \leq 1 - x + \frac{x^2}{2},\ \forall x \geq 0$$

  egyenlőtlenség is, amit a standard normális eloszlás sűrűségfüggvényének az

$$1 - \frac{x^2}{2} \leq \sqrt{2\pi} f_{\mathcal{N}(0,1)}(x) \leq 1 - \frac{x^2}{2} + \frac{x^4}{8},\ \forall x \geq 0 \qquad (5)$$

  alakú közrefogására használhatunk fel (az x paramétert az $\frac{x^2}{2}$ törttel helyettesítettük).

- Tekintsük a $c_1 = \sqrt{\frac{2\pi}{e}}$ optimális elutasítási konstanst és a $\theta = 1$ paraméterű Cauchy-féle $g_1(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$ sűrűségfüggvényt, amelyeket a 4. labor 5. algoritmusához szerkesztettünk!
- Tanulmányozzuk az 1. ábrát és a $W = \sqrt{2\pi} c_1 U g_1(Y)$ valószínűségi változót, ahol Y Cauchy-eloszlású $\theta = 1$ paraméterrel, U pedig egyenletes eloszlású a $[0,1]$ intervallumban!



1. ábra. Az (5)-ös egyenlőtlenségben, valamint a $W = \sqrt{2\pi} c_1 U g_1(Y)$ valószínűségi változóban szereplő függvények alakja.

- Az 1. ábrán feltüntetett függvények alakjából következik, hogy a
$$1-\frac{x^2}{2}+\frac{x^4}{8}\leq\sqrt{2\pi}c_1 g_1(x)$$
egyenlőtlenség nem teljesül minden $x\in\mathbb{R}$ esetén, így a közrefogási módszer által igényelt h_2 függvényt csak a
$$h_2(x)=\sqrt{2\pi}c_1 g_1(x),\ x\in\mathbb{R}$$
alakban választhatjuk meg, vagyis a közrefogás algoritmusból hiányozni fog a $W>h_2(Y)$ gyors elutasítási lépés.
- A $W\leq h_1(Y)$ gyors elfogadási lépéshez szükséges h_1 függvényt viszont megválaszthatjuk a
$$h_1(x)=1-\frac{x^2}{2}$$
alakban, amennyiben $|x|\leq\sqrt{2}$.
- Tehát a 4. labor 5. algoritmusához képest a közrefogás módszerével egy valamivel gyorsabb standard normális eloszlású számgenerátort implementálhatunk az 1. algoritmus alábbi módosított változatával.

2. Algoritmus (Közrefogás módszerén alapuló $\mathcal{N}(0,1)$-eloszlású számgenerátor)

- *Bemenet:* az $n \geq 1$ természetes szám, amely a kimeneti mintavétel méretét rögzíti.
- *Kimenet:* az $X \sim \mathcal{N}(0,1)$ valószínűségi változónak egy n-elemű független $\{X_i\}_{i=1}^{n}$ mintavétele.

1. $\alpha \leftarrow \dfrac{1}{\sqrt{e}}$, $\beta \leftarrow \dfrac{1}{2}$, $\gamma \leftarrow \sqrt{2}$
2. **Minden** $i = 1, 2, \ldots, n$ index esetén **végezd el**:
3.   **ismételd:**
4.   generáld a független $U \sim \mathcal{U}([0,1])$ és $V \sim \mathcal{U}([0,1])$ valószínűségi változókat
5. $Y \leftarrow \tan(\pi V)$
6. $S \leftarrow \beta Y^2$
7. $W \leftarrow \dfrac{\alpha U}{\beta + S}$

2. Algoritmus (folytatás)

ha $|Y| > \gamma$
  akkor $L \leftarrow$ hamis
  máskülönben $L \leftarrow [W \leq 1 - S]$
ha $[L ==$ hamis$]$, **akkor** $L \leftarrow \left[W \leq e^{-S}\right]$
ameddig $[L ==$ igaz$]$
$X_i \leftarrow Y$.

- Ekkor minden egyes mintavételi érték generálásakor a standard normális eloszlás $f_{\mathcal{N}(0,1)}$ sűrűségfüggvényét várhatóan

$$c_1 - \frac{1}{\sqrt{2\pi}} \int_{|x| \leq \sqrt{2}} \left(1 - \frac{x^2}{2}\right) \mathrm{d}x = \sqrt{\frac{2\pi}{e}} - \frac{4}{3\sqrt{\pi}} \approx 0,76809$$

alkalommal értékeljük ki.

- Megjegyezzük, hogy számos más véletlenszám-generátor is létezik, és ezek változó hatékonyságú algoritmusai más és más elméleti eredményekre épülhetnek.
- A mostani és az előző laborokon ismertetett inverziós, elutásítás és közrefogás módszererén alapuló algoritmusokkal csak egy rövid bepillantást szerettünk volna nyújtani a nemegyenletes eloszlású diszkrét és folytonos valószínűségi változók mintavételezésébe.
- Az érdeklődő hallgatók számára a [Devroye, 1986, W.H. Press et al., 2007] munkákat ajánlhatjuk.

Nemegyenletes eloszlású véletlenszám-generátorok

Leadási határidő: 2023. november 7–10. (a megfelelő laborórákon)

1. feladat

- Implementáljátok a közrefogás módszerét használó 2. algoritmust és teszteljétek, hogy hány lépés alatt fogadunk el egy mintát, illetve, hogy hányszor értékeljük ki a standard normális eloszlás sűrűségfüggvényét!

Opcionális feladat (+1 pont elméleti megoldással együtt)

- Határozzátok meg az $\alpha > 0$ paraméter értékét úgy, hogy az

$$f_X(x) = \begin{cases} \alpha \left[\frac{1}{4} + \sin^4\left(\frac{x}{3}\right)\right], & x \in [-3\pi, 3\pi], \\ 0, & x \notin [-3\pi, 3\pi] \end{cases}$$

  leképezés egy abszolút folytonos (kompakt tartományon értelmezett) valószínűségi változó sűrűségfüggvénye legyen, majd alkalmazzátok a közrefogás módszerét az adott valószínűségi változó mintavételezésére. Hasonlítsátok össze az így kapott mintavétel hisztogramát a korrigált sűrűségfüggvény alakjával!
- Az algoritmusotok várhatóan hányszor értékeli ki a mintavételezett valószínűségi változó sűrűségfüggvényét egy-egy minta generálásakor?

G. Marsaglia, **1977**.
The squeeze method for generating gamma variates,
Computers and Mathematics with Applications, **3**(4):321–325.

L. Devroye, **1986**.
Non-Uniform Random Variate Generation,
Springer-Verlag New York Inc.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, **2007**.
Numerical recipes. The art of scientific computing, 3rd Edition,
Cambridge University Press.