Hochschule RheinMain
Prof. Dr. D. Lehmann SS 2020

Week8: Monte Carlo Simulation

Es sei

$$p : \mathbb{R}^d \to \mathbb{R}$$

eine ein- oder mehrdimensionale Wahrscheinlichkeitsdichte, also $p \geq 0$, $d \geq 1$ und

$$\int_{\mathbb{R}^d} p(\phi)\, d^d\phi = 1 \, .$$

Für die meisten der gängigsten Wahrscheinlichkeitsverteilungen existieren in R sehr performante Zufallszahlgeneratoren, mit denen man also p-verteilte Zufallszahlen erzeugen kann. Das Kapitel 5 im RSkript_UniGiessen gibt dazu auf 3 Seiten eine sehr kompakte und verständliche Übersicht, das waren diese Sachen:

http://hsrm-mathematik.de/SS2020/semester4/Datenanalyse-und-ScientificComputing-mit-R/W'keitsverteilungen-in-R.pdf

Es sei jetzt $F : \mathbb{R}^d \to \mathbb{R}$ eine gegebene Funktion. Wir wollen das folgende Integral, das man auch als Erwartungswert von F interpretieren kann, berechnen[1]:

$$\mathsf{E}[F] \;=\; \int_{\mathbb{R}^d} F(\phi)\, p(\phi)\; d^d\phi$$

Dazu erzeugt man N p-verteilte, unabhängige Zufallszahlen, oder, falls $d > 1$, N Vektoren von Zufallszahlen

$$(\phi_1, \phi_2, \cdots, \phi_N) \;\in\; (\mathbb{R}^d)^N \;=\; \mathbb{R}^{dN}$$

und bildet die Monte Carlo Summe (beachten Sie, dass nur das F in der Monte Carlo Summe auftaucht, das p steckt schon in den Zufallszahlen drin)

$$S_N(F) \;:=\; \frac{1}{N}\sum_{i=1}^{N} F(\phi_i) \tag{1}$$

etwa mit $N = 10000$ oder $N = 100000$. Grundlage der Monte Carlo Berechnung ist dann die folgende approximative Identität:

$$S_N(F) \;\overset{N\to\infty}{\approx}\; \mathsf{E}[F] \;=\; \textstyle\int_{\mathbb{R}^d} F(\phi)\, p(\phi)\; d^d\phi \tag{2}$$

Bevor wir das ausprobieren, wollen wir uns den theoretischen Hintergrund dazu kurz anschauen. Dazu erinnern wir uns an den zentralen Grenzwertsatz aus dem `week4.txt`, wir nehmen die Version2:

[1] ist p eine diskrete Wahrscheinlichkeitsverteilung, so ist das Integral durch eine entsprechende Summe zu ersetzen

Theorem (Zentraler Grenzwertsatz): Es seien $X_1, X_2, X_3, \cdots$ eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit

$$\begin{aligned} \mathsf{E}[X_k] &= \mu \\ \mathsf{V}[X_k] &= \sigma^2 \end{aligned}$$

Wir betrachten die Mittelwerte (die nennen wir hier gleich S_n anstatt M_n, weil das sind dann genau die Mittelwerte aus Gleichung (1) oben, das grosse N ist hier ein kleines n)

$$S_n \;:=\; \tfrac{1}{n}\big(X_1 + X_2 + \cdots + X_n\big) \tag{3}$$

die den Erwartungswert

$$\mathsf{E}[S_n] \;=\; \mu \tag{4}$$

und die Varianz

$$\mathsf{V}[S_n] \;=\; \tfrac{\sigma^2}{n} \tag{5}$$

haben. Wir definieren die normierten Grössen

$$Z_n \;:=\; \tfrac{\sqrt{n}}{\sigma}\,(\,S_n \,-\, \mu\,) \tag{6}$$

Dann gilt

$$\lim_{n\to\infty} \mathsf{Prob}\big[\,Z_n \in (x, x+dx)\,\big] \;=\; \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx \quad . \tag{7}$$

Aus der Gleichung (7) folgt:

$$\begin{aligned} \mathsf{Prob}\big[\,|Z_n| \,\geq \alpha\,\big] \;&\overset{n\to\infty}{=}\; \int_{-\infty}^{-\alpha} \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx \;+\; \int_{\alpha}^{\infty} \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx \\ &=\; 2\int_{-\infty}^{-\alpha} \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx \;=\; 2\,\Phi(-\alpha) \end{aligned} \tag{8}$$

wobei die $\Phi(x)$-Funktion also definiert ist durch

$$\Phi(x) \;:=\; \int_{-\infty}^{x} \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{y^2}{2}}\, dy \quad . \tag{9}$$

Gleichung (8) können wir auch so schreiben:

$$\mathsf{Prob}\Big[\,\big|\tfrac{\sqrt{n}}{\sigma}\,(\,S_n \,-\, \mu\,)\big| \,\geq \alpha\,\Big] \;\overset{n\to\infty}{=}\; 2\,\Phi(-\alpha)$$

oder

$$\mathsf{Prob}\Big[\,\big|(\,S_n \,-\, \mu\,)\big| \,\geq \tfrac{\sigma}{\sqrt{n}}\,\alpha\,\Big] \;\overset{n\to\infty}{=}\; 2\,\Phi(-\alpha) \tag{10}$$

Als unabhängige, identisch verteilte Zufallsvariablen wählen wir jetzt also

$$X_i \;:=\; F(\phi_i) \tag{11}$$

so dass die S_n's dann also mit der Monte Carlo Summe ($n \to N$)

$$S_N \;=\; S_N(F) \;=\; \frac{1}{N}\sum_{i=1}^{N} F(\phi_i) \tag{12}$$

übereinstimmen. Wir haben dann

$$\mu \;=\; \mathsf{E}\big[\,F(\phi_i)\,\big] \;=\; \mathsf{E}\big[\,F(\phi)\,\big] \;=\; \int_{\mathbb{R}^d} F(\phi)\,p(\phi)\,d^d\phi \tag{13}$$

also das ist genau das Integral, was wir berechnen wollen, und die Grösse

$$\sigma^2 \;=\; \mathsf{V}[F] \;=\; \mathsf{E}[F^2]-\mathsf{E}[F]^2 \;=\; \int_{\mathbb{R}^d} F(\phi)^2 p(\phi)\,d^d\phi \;-\; \left(\int_{\mathbb{R}^d} F(\phi)\,p(\phi)\,d^d\phi\right)^2$$

hat dann einen Einfluss auf den Monte Carlo Error. Die Gleichung (10) lautet dann also

$$\mathsf{Prob}\Big[\;\big|\,S_N(F)-\mathsf{E}[F]\,\big| \;\geq\; \alpha\sqrt{\tfrac{\mathsf{V}[F]}{N}}\;\Big] \;\;\overset{N\to\infty}{=}\;\; 2\,\Phi(-\alpha) \tag{14}$$

Schauen wir uns Gleichung (14) genauer an. Das $\mathsf{E}[F]$ ist das exakte, das theoretische Resultat, was wir berechnen wollen. Das $S_N = S_N(f)$ ist die Monte Carlo Summe, diese Grösse tun wir also konkret numerisch berechnen indem wir Zufallszahlen ϕ_i simulieren und dann die Summe in (12) berechnen. Wir möchten dann natürlich, dass das S_N das $\mathsf{E}[F]$ möglichst gut 'trifft'. Da das S_N ja mit Zufallszahlen berechnet wird, kann es natürlich immer mal sein, dass die Zahlen jetzt gerade etwas komisch waren und die Sache nicht so gut hinkommt. Was wir fordern können, ist, dass die Wahrscheinlichkeit, dass so etwas passiert, möglichst klein ist.

Gleichung (14) sagt, dass man schonmal daneben liegen kann, aber die Wahrscheinlichkeit dafür ist kleiner als $2\Phi(-\alpha)$. Wir können dann also etwa fordern:

$$2\,\Phi(-\alpha) \;\overset{!}{<}\; 1\%\;, \tag{15}$$

wir liegen in höchstens 1% aller Fälle daneben, oder etwa

$$2\,\Phi(-\alpha) \;\overset{!}{<}\; 0.01\% \tag{16}$$

wir liegen in höchstens 0.01% aller Fälle daneben, und bekommen dann die folgenden Werte für das α:

$$2\,\Phi(-2.58) \;=\; 1\% \tag{17}$$

$$2\,\Phi(-3.89) \;=\; 0.01\% \tag{18}$$

Das heisst, mit einer Wahrscheinlichkeit von 99% ist

$$S_N(F)-2.58\sqrt{\tfrac{\mathsf{V}[F]}{N}} \;\leq\; \mathsf{E}[F] \;\leq\; S_N(F)+2.58\sqrt{\tfrac{\mathsf{V}[F]}{N}}$$

und mit einer Wahrscheinlichkeit von 99.99% ist

$$S_N(F)-3.89\sqrt{\tfrac{\mathsf{V}[F]}{N}} \;\leq\; \mathsf{E}[F] \;\leq\; S_N(F)+3.89\sqrt{\tfrac{\mathsf{V}[F]}{N}}$$

In dem `week8.txt` wollen wir jetzt die Monte Carlo Approximationsformel

$$\mathsf{E}[F] \;\approx\; \frac{1}{N}\sum_{i=1}^{N} F(\phi_i) \tag{19}$$

sowie die Abschätzung für den Fehler

$$\left|\, \mathsf{E}[F] \,-\, \frac{1}{N}\sum_{i=1}^{N} F(\phi_i) \,\right| \;\leq\; \frac{\mathrm{const}_d}{\sqrt{N}} \tag{20}$$

durch explizite Simulation an 3 Beispielen, zwei eindimensionale und ein mehrdimensionales, verdeutlichen.

Bsp 1) Überprüfen Sie, für $\lambda = 1$ und $\lambda = 2$, die Formel

$$\int_{-\infty}^{\infty} e^{\lambda x}\, e^{-\frac{x^2}{2}}\, \tfrac{dx}{\sqrt{2\pi}} \;=\; e^{\frac{\lambda^2}{2}} \tag{21}$$

mit Hilfe einer Monte Carlo Simulation, indem Sie normalverteilte Zufallszahlen verwenden. Was also genau ist F und was ist p? Plotten Sie die Monte Carlo Summe für

$$N \;\in\; \{1000, 2000, 3000, \cdots, 48000, 49000, 50000\} \tag{22}$$

Tragen Sie in diesen Plot ebenfalls das exakte Resultat als rote, horizontale Linie ein. Berechnen und plotten Sie dann, vielleicht nur für $\lambda = 2$, den Fehlerterm

$$\text{mc\_error} \;=\; \mathsf{E}[F] \,-\, \frac{1}{N}\sum_{i=1}^{N} F(\phi_i) \tag{23}$$

für die Werte von N aus (22). Plotten Sie ebenfalls die Grössen $\sqrt{N} \times \text{mc\_error}$ und $N \times \text{mc\_error}$ als Funktion von N.

Bsp 2) Überprüfen Sie die Formel

$$\int_0^{\infty} x^4\, e^{-x}\, dx \;=\; 4!$$

mit Hilfe einer Monte Carlo Simulation, indem Sie exponential-verteilte[2] Zufallszahlen verwenden. Was also genau ist F und was ist p? Plotten Sie die Monte Carlo Summe für die Werte von N aus (22). Tragen Sie in diesen Plot ebenfalls das exakte Resultat als rote, horizontale Linie ein. Berechnen und plotten Sie dann den Fehlerterm mc_error aus (23) für die Werte von N aus (22). Plotten Sie ebenfalls die Grössen $\sqrt{N} \times \text{mc\_error}$ und $N \times \text{mc\_error}$ als Funktion von N.

Bsp 3) Überprüfen Sie die Formel

$$\int_0^1 \int_0^1 \int_0^1 \chi(0 \leq x + y + z \leq 1)\, dx\, dy\, dz \;=\; \frac{1}{3!}$$

mit Hilfe einer Monte Carlo Simulation, indem Sie auf dem Intervall [0,1] gleichverteilte Zufallszahlen verwenden. Was also genau ist F und was ist p? Plotten Sie die Monte Carlo Summe für die Werte von N aus(22).

[2] schauen Sie sich dazu die Übersicht in `W'keitsverteilungen-in-R.pdf` an

Schliesslich: Auf dem letzten Übungsblatt hatten wir die Frage aufgeworfen, wie es sein kann, dass (mit den Notationen vom letzten Übungsblatt)

$$\pi_N \quad \approx \quad e^{-1.27 \times N}$$

und gleichzeitig

$$\mathsf{E}[\pi_N] \quad = \quad 1$$

gilt, das schauen wir uns jetzt auf dem neuen Übungsblatt an, eine sehr interessante Sache.