Metody Monte Carlo i MCMC

Maciej Romaniuk

Instytut Badań Systemowych PAN

email: mroman@ibspan.waw.pl

Warszawa, 29 lipca 2014

Spis treści

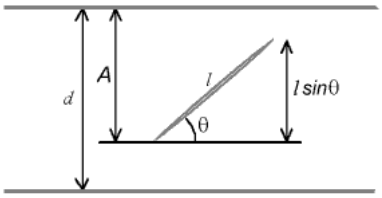
Rozdział 1

Wprowadzenie

1.1 Przykłady zastosowań

Przykład 1.1 (Igła Buffona)**.** *Załóżmy, że na drewnianą podłogę złożoną z desek o szerokości d rzucamy igłę o długości l. Jakie jest prawdopodobieństwo, że rzucona igła przetnie choć jedną z linii pomiędzy deskami?*

Rozwiązanie: Niech A będzie odległością dolnego końca igły od wybranej linii na deskach, a Θ – kątem nachylenia igły względem linii (patrz rysunek 1.1). Rozkład prawdopodobieństwa rzucenia igły na podłogę jest wtedy rozkładem jednostajnym na przestrzeni $A \times \Theta = [0; d) \times [0; \pi)$.



Rysunek 1.1: Problem Igły Buffona

Jak łatwo zauważyć, prawdopodobieństwo przecięcia jednej z linii przez upadającą igłę wynosi

$$\int_0^{\pi} \int_0^{l \sin \Theta} \frac{1}{d\pi} \, dA \, d\Theta = \frac{2l}{d\pi} \, . \tag{1.1}$$

Co najciekawsze, wzór (1.1) można łatwo wykorzystać w zupełnie innym celu – do estymacji, czyli znalezienia przybliżenia, wartości liczby π.

Niech n będzie mianowicie liczbą wszystkich prób, a k – liczbą rzutów, w których igła przecięła linię na deskach. Wtedy

$$\hat{\pi} = \frac{2ln}{dk} \, , \tag{1.2}$$

gdzie $\hat{\pi}$ jest tzw. estymatorem Monte Carlo szukanej wartości liczby π. $\diamond$

Przykład 1.2 (Pole skomplikowanego kształtu)**.** *Załóżmy, że mamy pewien zamknięty kształt zawarty w kwadracie o wymiarach metr na metr. Jak można zmierzyć pole tego (nawet skomplikowanego) kształtu?*

Rozwiązanie: Należy upuścić (w losowy sposób) na kwadrat odpowiednio dużą liczbę punktów. Jeśli przez n oznaczymy liczbę wszystkich wykorzystanych punktów, a przez k liczbę punktów, które znalazły się w środku kształtu, to otrzymujemy przybliżenie (dokładniej – estymator metody Monte Carlo) szukanego pola figury w postaci

$$\text{Pole} \approx \frac{k}{n} \tag{1.3}$$

◇

Przykład 1.3 (Obliczanie całek)**.** *Załóżmy, że chcemy obliczyć całkę dla pewnej skomplikowanej funkcji $f(x)$ postaci*

$$\int_0^1 f(x)dx \ . \tag{1.4}$$

Jak możemy to zrobić korzystając z losowania?

Rozwiązanie: Załóżmy, że dysponujemy próbą $X_1, X_2, \ldots, X_n$ punktów „wylosowanych równomiernie" (czyli pochodzących z rozkładu jednostajnego) na odcinku $[0; 1]$. Wtedy wartość

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{1.5}$$

będzie estymatorem (metody Monte Carlo) i dla dostatecznie dużych wartości n będzie odpowiednio dobrze przybliżać szukaną całkę. ◇

Przykład 1.4 (Niezawodność sieci opisanej grafem)**.** *Dysponujemy siecią (elektryczną, informatyczną, itp.) opisaną spójnym grafem. Wierzchołki opisują użytkowników sieci, krawędzie – odpowiednie linie (energetyczne, przesyłowe, itp.). Krawędzi jest w sumie m. Załóżmy, że krawędź i-ta działa z prawdopodobieństwem p_i, a nie działa z prawdopodobieństwem $1 - p_i$. Jeśli krawędzie mogą się psuć niezależnie od siebie, to jakie jest prawdopodobieństwo działania całej sieci jednocześnie?*

Rozwiązanie: Ponieważ krawędzie (linie) psują się niezależnie od siebie, to prawdopodobieństwo niezawodności dla całej sieci wynosi

$$\mathrm{P} = \prod_{i=1}^{m} p_i \ . \tag{1.6}$$

Prawdopodobieństwo to możemy otrzymać również drogą odpowiedniego losowania. W tym celu tworzymy n „realizacji" całej sieci. Dla każdej z tych realizacji losujemy, czy poszczególne krawędzie w sieci działają, czy nie. Następnie obliczamy frakcję wszystkich w pełni działających realizacji w wygenerowanej puli n symulacji. ◇

Przykład 1.5 (Minimum funkcji)**.** *Chcemy znaleźć minimum skomplikowanej funkcji $f(x)$, która zdefiniowana jest na przedziale $[0; 1]$. Jak możemy to zrobić?*

Rozwiązanie: Wystarczy w tym celu wygenerować próbę $X_1, X_2, \ldots, X_n$ punktów „wylosowanych równomiernie" (czyli pochodzących z rozkładu jednostajnego) na odcinku $[0; 1]$. Wtedy

$$\min\{f(X_1), f(X_2), \ldots, f(X_n)\} \tag{1.7}$$

będzie przybliżeniem (estymatorem metody Monte Carlo) dla minimum funkcji $f(x)$. ◇

1.2 Podstawowe definicje i twierdzenia

1.2.1 Prawdopodobieństwo

- **Doświadczenie losowe (eksperyment losowy)** – doświadczenie, którego wyniku nie jesteśmy w stanie przewidzieć i które jest modelowane matematycznie przy pomocy zasad rachunku prawdopodobieństwa (np. rzut monetą, wybór kuli z urny, czas do przyjazdu autobusu).

- **Rezultat doświadczenia losowego** – wynik doświadczenia losowego, np. wynik rzutu monetą (orzeł lub reszka), wybrana kula z urny (biała, czarna, czerwona).

- **Przestrzeń zdarzeń elementarnych** – zbiór wszystkich możliwych wyników (rezultatów) doświadczenia, złożony ze **zdarzeń elementarnych**, czyli najprostszych, nierozkładalnych rezultatów danego eksperymentu losowego. Zazwyczaj oznaczany jest przez Ω.

- **Zdarzenie losowe** – dowolny podzbiór przestrzeni zdarzeń elementarnych, zazwyczaj oznaczany symbolem ω.

Uwaga! Należy pamiętać, że zdarzenie losowe, które rozpatrujemy, może być „dowolne", ale jednocześnie musi być „dostatecznie porządne". Wiąże się to ściśle z teorią miary w matematyce, której tutaj nie będziemy rozwijać. W praktycznych zastosowaniach właściwie wszystkie zdarzenia losowe, o prawdopodobieństwie których będziemy mówili są „dostatecznie porządne". Dokładniej rzecz ujmując, rodzinę tych „odpowiednich" zdarzeń losowych będziemy oznaczali przez $\mathcal{F}$ i rozumieli przez nią pewne σ-ciało (np. zbiorów otwartych w Ω).

1.2.2 Prawdopodobieństwo warunkowe

Definicja 1.6. *Niech A, B będą zdarzeniami losowymi i $\mathrm{P}(B) > 0$.* **Prawdopodobieństwem warunkowym** *zdarzenia A pod warunkiem zdarzenia B nazywać będziemy*

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \ . \tag{1.8}$$

Intuicyjnie biorąc, wiedza, że zaszło zdarzenie B może zmieniać naszą ocenę co do szans zajścia zdarzenia A. Ta dodatkowa informacja jak gdyby „zawęża" przestrzeń probabilistyczną.

1.2.3 Prawdopodobieństwo całkowite. Wzór Bayesa

Twierdzenie 1.7 (Prawdopodobieństwo całkowite)**.** *Niech zdarzenia* $A_1, A_2, \ldots \in \mathcal{F}$ *spełniają warunki:*

1. $A_i \cap A_j = \emptyset$ *dla* $i \neq j$
2. $A_1 \cup A_2 \cup \ldots = \Omega$
3. $\mathrm{P}(A_i) > 0$ *dla każdego* $i = 1, 2, \ldots$

Wtedy dla dowolnego zdarzenia A *zachodzi*

$$\mathrm{P}(A_i|A) = \sum_{i=1} \mathrm{P}(A|A_i)\,\mathrm{P}(A_i)\ . \tag{1.9}$$

Twierdzenie 1.8 (Wzór Bayesa)**.** *Przy założeniach z tw. o prawdopodobieństwie całkowitym i jeśli ponadto* $\mathrm{P}(A) > 0$*, to*

$$\mathrm{P}(A) = \frac{\mathrm{P}(A|A_i)\,\mathrm{P}(A_i)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|A_i)\,\mathrm{P}(A_i)}{\sum_{i=1} \mathrm{P}(A|A_i)\,\mathrm{P}(A_i)}\ . \tag{1.10}$$

Wzór Bayesa umożliwia „odwrócenie" pytania, tzn. poszukiwanie prawdopodobieństwa „przyczyny" przy znanym prawdopodobieństwie dotyczącym „skutku".

1.2.4 Niezależność zdarzeń

Definicja 1.9. *Zdarzenia* A *i* B *nazwiemy* **niezależnymi***, jeżeli*

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)\ . \tag{1.11}$$

Niezależność probabilistyczna zdarzeń ma wiele wspólnego z intuicyjnie rozumianym „brakiem wpływu". Jeśli A i B są niezależne oraz $\mathrm{P}(B) > 0$, to

$$\mathrm{P}(A|B) = \mathrm{P}(A)\ . \tag{1.12}$$

Niezależność zdarzeń możemy rozszerzyć na więcej niż tylko dwa zdarzenia, wymaga to jednak sprawdzenia większej liczby warunków.

Definicja 1.10. *Zdarzenia* $A_1, A_2, A_3, \ldots, A_n \in \mathcal{F}$ *są* **niezależne***, jeśli spełniony jest układ równań*

$$\mathrm{P}(A_i \cap A_j) = \mathrm{P}(A_i)\,\mathrm{P}(A_j)\ (i < j)\ , \tag{1.13}$$

$$\mathrm{P}(A_i \cap A_j \cap A_k) = \mathrm{P}(A_i)\,\mathrm{P}(A_j)\,\mathrm{P}(A_k)\ (i < j < k)\ , \ldots \tag{1.14}$$

$$\mathrm{P}(A_1 \cap \ldots \cap A_n) = \mathrm{P}(A_1) \ldots \mathrm{P}(A_n)\ . \tag{1.15}$$

1.2.5 Zmienna losowa

Definicja 1.11. *Funkcję X określoną na przestrzeni zdarzeń elementarnych Ω o wartościach w przestrzeni rzeczywistej nazywamy* **zmienną losową**, *jeśli dla każdego $t \in \mathbb{R}$ zbiór*

$$\{\omega \in \Omega : X(\omega) \leqslant t\} \tag{1.16}$$

jest zdarzeniem losowym (czyli należy do $\mathcal{F}$).

Pojęcie zmiennej losowej jest bardzo użyteczne. Po pierwsze, w większości zastosowań rzeczywiście interesują nas pewne wartości liczbowe związane z modelem probabilistycznym, np. ilość oczek na wyrzuconej ściance kości, czas do przyjazdu autobusu, czy do momentu pierwszej awarii urządzania. Po drugie, wykorzystanie tego pojęcia pozwala nam się „oderwać" od (być może bardzo skomplikowanej) przestrzeni probabilistycznej Ω i przejść do „bardziej naturalnej" dziedziny liczb rzeczywistych, nie tracąc przy tym nic w opisie modelu.

Wykorzystany w definicji warunek ma znaczenie głównie techniczne – dzięki temu zapewniamy funkcji X mierzalność względem σ-ciała $\mathcal{F}$. W praktyce każda dostatecznie „rozsądna" funkcja określona na przestrzeni Ω jest zmienną losową.

Definicja 1.12. Rozkład prawdopodobieństwa zmiennej losowej *X jest to funkcja, która zbiorowi $B \subset \mathbb{R}$ przyporządkowuje liczbę*

$$\mathrm{P}(X \in B) = \mathrm{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right) \ . \tag{1.17}$$

Definicja 1.13. Dystrybuanta *zmiennej losowej X jest to funkcja $F : \mathbb{R} \to \mathbb{R}$ określona wzorem*

$$F(a) = \mathrm{P}(X \leqslant a) = \mathrm{P}\left(\{\omega \in \Omega : X(\omega) \leqslant a\}\right) \ . \tag{1.18}$$

1.2.6 Rozkład dyskretny zmiennej losowej

Definicja 1.14. *Zmienna losowa X ma* **rozkład dyskretny**, *jeśli zbiór jej wartości jest skończony lub przeliczalny (czyli równoliczny ze zbiorem liczb naturalnych). Zatem dla pewnego zbioru $\mathcal{X} = \{x_1, x_2, \ldots\}$ mamy $\mathrm{P}(X \in \mathcal{X}) = 1$.*

Zmienną losową X można w bardzo prosty sposób opisać. Wystarczy podać jakie prawdopodobieństwo jest przyjmowane dla każdej wartości $x_i \in \mathcal{X}$, czyli (np. w postaci formuły lub tabelki) określić

$$\mathrm{P}(X = x_i) = p_i = f(x_i) \ . \tag{1.19}$$

Dystrybuanta takiej zmiennej losowej ma charakterystyczną, „schodkową" postać i jest określona wzorem

$$F_X(a) = \mathrm{P}(X \leqslant a) = \sum_{x_i \leqslant a} p_i = \sum_{x_i \leqslant a} f(x_i) \ . \tag{1.20}$$

1.2.7 Rozkład ciągły zmiennej losowej

Definicja 1.15. *Jeśli istnieje nieujemna funkcja $f : \mathbb{R} \to \mathbb{R}$ t.ż. dla dowolnych $a < b$ mamy*

$$\mathrm{P}(a < X < b) = \mathrm{P}(a \leqslant X \leqslant b) = \int_a^b f(x)dx \tag{1.21}$$

to zmienna X ma rozkład **ciągły** *z* **gęstością prawdopodobieństwa** *określoną funkcją $f(x)$.*

Uwaga! Ściśle rzecz biorąc, rozkład spełniający powyższą definicję nazywany jest absolutnie ciągłym, a mówimy, że rozkład jest ciągły jeśli jego dystrybuanta jest ciągła. Istnieją bowiem ciągłe dystrybuanty bez gęstości (czyli nie-absolutnie ciągłe), ale nie będziemy się nimi tutaj zajmowali.

1.2.8 Momenty zmiennych losowych

Definicja 1.16. Wartością oczekiwaną *(czasami też określaną jako średnią) zmiennej losowej X nazywamy liczbę $\mathbb{E}\,X$ określoną wzorem*

$$\mathbb{E}\,X = \sum_{i=1} x_i p_i \tag{1.22}$$

dla zmiennej losowej o rozkładzie dyskretnym, lub

$$\mathbb{E}\,X = \int_{\mathbb{R}} x f(x)\,dx \tag{1.23}$$

dla zmiennej losowej o rozkładzie ciągłym określonym funkcją gęstości $f(x)$, przy założeniu, że odpowiedni szereg lub całka są bezwzględnie zbieżne.

Często niezbędne jest obliczenie wartości oczekiwanej pewnej funkcji $g(.)$ określonej jako funkcja zmiennej losowej X. Mamy wtedy odpowiednio

$$\mathbb{E}\,g(X) = \sum_{i=1} g(x_i) p_i \tag{1.24}$$

lub

$$\mathbb{E}\,g(X) = \int_{\mathbb{R}} g(x) f(x)\,dx \tag{1.25}$$

znowu przy istotnym założeniu bezwzględnej zbieżności wyrażeń po prawej stronie wzorów. Ponieważ założenie to będzie się cały czas powtarzać w kolejnych definicjach, nie będziemy już go przywoływać.

Definicja 1.17. Wariancją *zmiennej losowej X nazywamy*

$$\operatorname{Var} X = \mathrm{D}^2\,X = \mathbb{E}(X - \mathbb{E}\,X)^2\ . \tag{1.26}$$

Definicja 1.18. Odchyleniem standardowym *zmiennej losowej X nazywamy*

$$\mathrm{D}\,X = \sqrt{\mathbb{E}(X - \mathbb{E}\,X)^2}\ . \tag{1.27}$$

Powyższe definicje możemy uogólnić.

Definicja 1.19. Momentem zwykłym rzędu *r zmiennej losowej X nazywamy $\mathbb{E}\,X^r$, zaś* **momentem centralnym rzędu** *r nazywamy $\mathbb{E}(X - \mathbb{E}\,X)^r$.*

W ten sposób wartość oczekiwana jest momentem zwykłym pierwszego rzędu, a wariancja – momentem centralnym drugiego rzędu.

1.2.9 Wektory losowe

Definicja 1.20. *Niech X i Y będą zmiennymi losowymi określonymi na tej samej przestrzeni probablistycznej.* **Rozkład wektora losowego** (X, Y) *(lub inaczej* **łączny rozkład prawdopodobieństwa** (X, Y)*) jest to funkcja, która zbiorowi $A \subset \mathbb{R}^2$ przyporządkowuje liczbę*

$$\mathrm{P}((X,Y) \in B) = \mathrm{P}\left(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in B\}\right) \ . \tag{1.28}$$

Łączny rozkład wektora losowego (X, Y) w prosty sposób definiuje rozkład pojedynczej zmiennej losowej (np. X), zwany wtedy **rozkładem brzegowym**

$$\mathrm{P}(X \in B) = \mathrm{P}((X,Y) \in B \times \mathbb{R}) \ . \tag{1.29}$$

Definicja 1.21. Dystrybuanta łącznego rozkładu prawdopodobieństwa *zmiennych losowych X i Y jest to funkcja $F : \mathbb{R}^2 \to \mathbb{R}$ dana wzorem*

$$F(a,b) = \mathrm{P}(X \leqslant a, Y \leqslant b) \ . \tag{1.30}$$

Gdyby prowadziło to do nieporozumień, dystrybuantę łącznego rozkładu prawdopodobieństwa zmiennych losowych X i Y oznacza się symbolem $F_{X,Y}$.

Definicja 1.22. *Łączny rozkład prawdopodobieństwa zmiennych losowych X i Y jest* **ciągły**, *jeśli dla dowolnego $B \in \mathbb{R}^2$ jest postaci*

$$\mathrm{P}((X,Y) \in B) = \int\int_B f(x,y)\, dx\, dy \tag{1.31}$$

dla pewnej nieujemnej funkcji $f(.,.)$ zwanej **łączną gęstością prawdopodobieństwa**.

Często stosujemy symbol $f_{X,Y}(.,.)$, aby odróżnić łączną gęstość prawdopodobieństwa od gęstości rozkładów brzegowych zmiennych X i Y, zwanych wtedy **gęstościami brzegowymi** $f_X(.)$ i $f_Y(.)$.

Definicja 1.23. *Zmienne losowe X i Y są* **niezależne**, *jeśli dla dowolnych zdarzeń $A \subset \mathbb{R}$ i $B \subset \mathbb{R}$ zachodzi*

$$\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\,\mathrm{P}(Y \in B) \ . \tag{1.32}$$

1.2.10 Model statystyczny

Statystyka matematyczna opiera się na założeniu, że obserwowane przez nas dane są wynikiem działania pewnego „mechanizmu losowego". Naszym celem w statystyce matematycznej jest właśnie poznanie i zidentyfikowania tego mechanizmu losowego.

Zakładać będziemy, że w naszym doświadczeniu losowym mamy do czynienia ze zmiennymi losowymi $X_1, X_2, \ldots, X_n$ określonymi na pewnej przestrzeni probabilistycznej, a obserwacje są realizacjami (wartościami, wynikami) tych zmiennych losowych. Zgodnie z ogólną konwencją owe zmienne losowe określone przez poszukiwany przez nas mechanizm losowy oznaczać będziemy dużymi literami, zaś zaobserwowane dane (wyniki) – literami małymi. W ten sposób $x_1 = X_1(\omega)$ dla pewnego $\omega \in \Omega$.

W statystyce przyjmujemy, że *nie znamy* rozkładu prawdopodobieństwa dla przestrzeni Ω, który „rządzi" zachowaniem zmiennych losowych $X_1, X_2, \ldots, X_n$. Chcemy go poznać i zidentyfikować na podstawie naszych danych – czyli obserwacji $x_1, x_2, \ldots, x_n$.

Definicja 1.24. Próbką *(lub* **próbą***) z rozkładu prawdopodobieństwa o dystrybuancie F nazywamy ciąg niezależnych zmiennych losowych $X_1, X_2, \ldots, X_n$, takich, że rozkład każdej zmiennej losowej X_i jest określony przez dystrybuantę F dla $i = 1, 2, \ldots, n$.*

Dla próby będziemy używać oznaczenia

$$X_1, X_2, \ldots, X_n \sim F \tag{1.33}$$

lub

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F \ , \tag{1.34}$$

gdzie **iid** jest skrótem od *independent, identically distributed* (niezależne, o jednakowym rozkładzie). Jeśli założymy, że zmienna X ma już rozkład określony dystrubuantą F, to czasami nadużywa się powyższych oznaczeń i zamiast (1.33) stosuje

$$X_1, X_2, \ldots, X_n \sim X \ . \tag{1.35}$$

W podobny sposób, jeśli założymy, że ciąg zmiennych losowych jest próbką z rozkładu normalnego, stosować będziemy skrócony zapis

$$X_1, X_2, \ldots, X_n \sim N(\mu; \sigma^2) \ , \tag{1.36}$$

zamiast „w pełni poprawnego"

$$X_1, X_2, \ldots, X_n \sim F_{N(\mu;\sigma^2)} \ . \tag{1.37}$$

Definicja 1.25. *Niech $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$. Funkcję*

$$\hat{F}(x) = \frac{1}{n} \sum_{i=0}^{n} \mathbb{1}(X_i \leqslant x) \tag{1.38}$$

nazywamy **dystrybuantą empiryczną**.

Czasami stosujemy zapis $\hat{F}_n(x)$ celem podkreślenia, że dystrybuanta empiryczna została skonstruowana w oparciu o n obserwacji. Dystrybuantę empiryczną traktujemy jako empiryczny, „obserwowalny" odpowiednik nieznanej dla nas dystrybuanty $F(x)$, która „odpowiada" za zgromadzone przez nas dane.

Twierdzenie 1.26 (Gliwienki – Cantellego)**.** *Jeżeli $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, to*

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F(x) \right| \xrightarrow[n \to \infty]{p.n.} 0 \ , \tag{1.39}$$

gdzie zbieżność następuje prawie na pewno*, tzn. dla każdego $\epsilon > 0$ zachodzi*

$$\sup_{x \in \mathbb{R}} \lim_{n \to \infty} \mathrm{P}\left(\left| \hat{F}_n(x) - F(x) \right| \leqslant \epsilon \right) = 1 \ . \tag{1.40}$$

Twierdzenie to oznacza, że wraz ze zwiększaniem się wielkości próbki (liczby obserwacji), możemy poznać (przybliżyć) nieznany nam rozkład prawdopodobieństwa w „mechanizmie losowym" z dowolną zadaną przez nas dokładnością.

W wielu przypadkach, gdy mowa o nieznanym „mechanizmie losowym", zakładamy jednak znajomość jego pewnych własności. Na przykład przy kontroli

elementów w fabryce, za rozsądne uznamy, że istnieją dwie możliwe wartości procesu kontrolnego – element będzie albo prawidłowy, albo nieprawidłowy. Możemy również założyć, że istnieje pewne prawdopodobieństwo zdarzenia, że element jest nieprawidłowy i wynosi ono np. θ. Niestety, dalej nie wiemy, ile owo prawdopodobieństwo θ wynosi, a co więcej – tak naprawdę dzięki statystyce chcemy owo θ znaleźć!

Definicja 1.27. Model statystyczny *określamy przez podanie przestrzeni* Ω*, rodziny rozkładów prawdopodobieństwa* $\{\mathrm{P}_\theta : \theta \in \Theta\}$ *indeksowanych parametrem* θ *oraz ciągu zmiennych losowych* $X_1, X_2, \ldots, X_n$ *zwanych obserwacjami.*

Rodzina rozkładów prawdopodobieństwa jest doprecyzowanym przez nas „mechanizmem losowym". Parametr θ odgrywa rolę etykiety identyfikującej poszczególne rozkłady prawdopodobieństwa. Najogólniej rzecz ujmując, naszym celem w statystyce jest poznanie wartości parametru θ. Wiemy bowiem, że nasze obserwacje są wynikiem działania pewnego rozkładu P_{θ_0}, ale nie wiemy, jakie jest to szczególne θ_0 w naszym przypadku.

Uwaga! Stosowanie **parametru** θ i **przestrzeni parametrów** Θ jest bardzo wygodnym sposobem opisu rzeczywistości. Parametr ten może być bowiem liczbą (np. $\theta \in [0; 1]$, jeśli rozważamy prawdopodobieństwo napotkania elementu nieprawidłowego przy kontroli jakości), jak i czymś znacznie bardziej skomplikowanym.

Uwaga! Bardzo często stosuje się oznaczenia (np. dystrybuanty) podkreślające związek z parametrem θ, np.

$$F_\theta(x) = \mathrm{P}_\theta(X \leqslant x) \ . \tag{1.41}$$

Definicja 1.28. Statystyką *nazywamy dowolną funkcję* T*, której argumentami są obserwacje, czyli*

$$T(X_1, X_2, \ldots, X_n) \ . \tag{1.42}$$

Uwaga! Dokładniej rzecz biorąc, $T : \mathbb{X}^n \to \mathbb{R}$, gdzie $X_i : \Omega \to \mathbb{X}$ (czyli $\mathbb{X}$ jest zbiorem wartości pojedynczej obserwacji, pojedynczej zmiennej losowej). Tak więc statystyka jest właściwie tym samym, co zmienna losowa. Jednak w znaczeniu różnych funkcji bazujących na obserwacjach łatwiej (i bardziej jednoznacznie) jest mówić o „statystyce" niż o zmiennej losowej.

1.2.11 Testy statystyczne

Testy statystyczne ogólnie rzecz biorąc polegają na sprawdzeniu poprawności jakiegoś zdania dotyczącego modelu statystycznego. Jeśli Θ jest przestrzenią parametrów modelu statystycznego, to możemy być zainteresowani prawdziwością następującego stwierdzenia „na podstawie obserwacji stwierdzamy, że wartość parametru θ wynosi dokładnie pięć". Innymi słowy, wyrażamy pewną opinię dotyczącą rozkładu „rządzącego" modelem statystycznym i uwzględniając zgromadzone dane, uznajemy tą opinię za prawdziwą (przyjmujemy naszą hipotezę) lub za fałszywą (odrzucamy hipotezę).

Owa opinia nazywana jest **hipotezą zerową** (oznaczana bywa zazwyczaj jako H_0). Statystycznie utożsamiamy ją z pewnym podzbiorem parametrów modelu $\Theta_0 \subset \Theta$ i zapisujemy w postaci

$$H_0 : \theta \in \Theta_0 \ , \tag{1.43}$$

czyli „czy jest prawdą, że nieznany nam parametr modelu θ pochodzi z pewnego ustalonego podzbioru parametrów Θ_0?". Obok hipotezy zerowej istnieje również **hipoteza alternatywna** (oznaczana jako H_1 lub K). Formułujemy ją w sposób następujący

$$H_1 : \theta \in \Theta_1 \ , \tag{1.44}$$

gdzie również $\Theta_1 \subset \Theta$, przy czym $\Theta_0 \cap \Theta_1 = \emptyset$. Oznacz to, że H_0 i H_1 są względem siebie konkurencyjne (albo – albo). Ponadto zazwyczaj $\Theta_1 = \Theta - \Theta_0$, co oznacza „jeśli H_0 nie jest prawdziwe, to musisz wybrać H_1, czyli zanegowanie H_0".

Trzeba pamiętać o tym, że hipotezy nie są *równoprawne*. „Bardziej" jesteśmy zainteresowani H_0 i będziemy się jej trzymać, o ile coś nas bardzo mocno nie przekona, że H_0 nie jest prawdą. O tym „czymś" mówi właśnie zasada konstruowania testu statystycznego.

Ogólna zasada testowania hipotez naukowych jest następująca: odrzuć hipotezę, jeśli rezultaty doświadczeń przeczą przyjętej hipotezie. Ponieważ w statystyce mamy do czynienia z losowością zjawisk (zmiennymi losowymi), *odrzucimy hipotezę zerową, jeśli przy założeniu jej prawdziwości otrzymane wyniki doświadczenia są bardzo mało prawdopodobne*. Np. jeśli rzucamy monetą tysiąc razy, to będziemy skłonni wątpić w jej symetryczność jeśli wypadną same orły (choć przecież wypadnięcie samych orłów przy rzucie symetryczną monetą tysiąc razy jest *możliwe* – tylko, że *bardzo mało prawdopodobne*). Wybieramy zatem pewne zdarzenie K, takie, że przy założeniu prawdziwości hipotezy zerowej $\mathrm{P}(K) = \alpha$ dla pewnego małego (bliskiego zera) α (np. $\alpha = 0,05$). K nazywamy obszarem krytycznym, a α – poziomem istotności testu. Procedura taka nosi czasami nazwę *testu zgodności*. Ponadto, w celu podkreślenia, że prawdopodobieństwo obserwacji obliczamy przy założeniu prawdziwości hipotezy zerowej, będziemy używać oznaczenia P_{H_0} lub P_{Θ_0}.

Przykład 1.29. *Rzuciliśmy monetą $n = 1000$ razy, przy czym otrzymaliśmy $k = 458$ orłów. Czy moneta jest symetryczna?*

Rozwiązanie: Łatwo zauważyć, że wielokrotne rzuty monetą możemy opisać rozkładem dwumianowym. Zatem, jeśli przez X oznaczymy liczbę otrzymanych orłów, mamy $X \sim \mathrm{Bin}(n; p)$. Pytanie o symetrię monety jest tym samym co postawienie hipotezy zerowej o treści

$$H_0 : p = \frac{1}{2} \ , \tag{1.45}$$

gdzie p jest prawdopodobieństwem otrzymania orła w pojedynczym rzucie.

Logicznym jest oczekiwanie (przy założeniu symetrii monety), że liczba wyrzuconych orłów powinna wynosić w przybliżeniu 500 ($n \cdot p$). Z kolei „zbyt mała" lub „zbyt duża" liczba orłów będzie przesłanką do odrzucenia hipotezy o symetrii naszej monety. Ponieważ nie wiemy, co znaczy dokładnie „zbyt mała" lub „zbyt duża" (100? 200? 300? 800? 900?) liczba orłów, musimy odwołać się do statystyki. Mamy zatem następującą regułę:

- jeśli $|X - 500| \leqslant c$ (dla pewnego c) to pozostajemy przy H_0 (czyli liczba orłów jest „odpowiednio bliska" 500)
- jeśli $|X - 500| > c$ to odrzucamy H_0 (czyli liczba orłów jest albo „zbyt mała", albo „zbyt duża").

Pozostaje zatem wyznaczyć wartość c, która na zadanym poziomie istotności α będzie spełniać wspomnianą wcześniej regułę, tzn.

$$\mathrm{P}_{H_0}(|X - 500| > c) = \alpha \ , \tag{1.46}$$

przy czym niech np. $\alpha = 0,05$.

Korzystając z centralnego twierdzenia granicznego, możemy przybliżyć skomplikowany dwumianowy rozkład zmiennej X prostszym obliczeniowo rozkładem normalnym. W ten sposób mamy $X \sim N(np; np(1-p))$ (w naszym przypadku $X \sim N(500; 250)$). Po zastosowaniu twierdzenia o standaryzacji mamy zatem

$$\mathrm{P}_{H_0}\left(\frac{|X - 500|}{\sqrt{250}} > \frac{c}{\sqrt{250}}\right) = \alpha \ , \tag{1.47}$$

co dla Y z rozkładu normalnego standardowego oznacza

$$\mathrm{P}\left(Y > \frac{c}{\sqrt{250}}\right) = \frac{\alpha}{2} = 0,025 \ , \tag{1.48}$$

dzięki czemu z tablic rozkładu normalnego standardowego otrzymujemy

$$\frac{c}{\sqrt{250}} = 1,96 \ , \tag{1.49}$$

co po zaokrągleniu do liczby całkowitej daje $c = 31$. Zatem

$$\mathrm{P}_{H_0}(|X - 500| > 31) \approx \alpha \ . \tag{1.50}$$

Otrzymaliśmy w ten sposób regułę testującą: jeśli liczba wyrzuconych oczek jest zbyt mała (mniejsza lub równa 478) lub zbyt duża (większa lub równa od 532) powinniśmy odrzucić hipotezę o symetrii monety. Dla $k = 458$ otrzymanego w naszym doświadczeniu hipotezę o symetrii monety trzeba więc odrzucić.

W rozważanym przez nas przypadku obszar krytyczny K ma postać

$$K = \{0, 1, \ldots, 468\} \cup \{532, 533, \ldots, 1000\} \ . \tag{1.51}$$

$\diamond$

Łatwo zauważyć, że w procedurze testowej posługiwaliśmy się pewną funkcją obserwacji (liczbą orłów w doświadczeniu losowym). Jest to zatem *statystyka*, a ponieważ wykorzystujemy ją do testowania, odpowiednią funkcję zmiennych $T(X_1, X_2, \ldots, X_n)$ w każdym z testów nazywamy **statystyką testową**

1.2.12 Estymacja

Zgodnie z wprowadzoną wcześniej definicją modelu statystycznego, po wybraniu rodziny rozkładów prawdopodobieństwa P_θ niezbędne jest wskazanie, przynajmniej w *przybliżeniu*, odpowiedniej wartości θ_0. Bazujemy przy tym na wynikach doświadczenia losowego, czyli naszych obserwacjach $X_1, \ldots, X_n$.

Definicja 1.30. Estymatorem *nieznanego parametru θ nazywamy dowolną statystykę $T(X_1, \ldots, X_n)$ o wartościach w zbiorze Θ.*

Estymator jest więc próbą zgadnięcia na podstawie obserwacji, ile wynosi wartość poszukiwanego parametru. Oczywiście chcielibyśmy, aby skonstruowany przez nas estymator odpowiednio *dobrze* przybliżał parametr θ.

Zazwyczaj estymator parametru θ oznaczamy przez dodanie „daszka", tzn. $\hat{\theta}$ w tym przypadku. Ogólniej, jeśli $g : \Theta \to \mathbb{R}$ jest pewną znaną funkcją parametru, to estymator będziemy oznaczali przez $\hat{g}$, ponieważ chcemy znaleźć przybliżenie $g(\theta)$.

Metoda największej wiarygodności w esytmacji polega na wybraniu takiej wartości parametru θ, który w danym przypadku jest dla nas „najbardziej wiarygodny", tzn. dla ustalonych wartości obserwacji jest „najbardziej prawdopodobny".

Niech $f_\theta(x_1, x_2, \ldots, x_n)$ oznacza łączną gęstość obserwacji (tzn. jeśli zmienne $X_1, X_2, \ldots, X_n$ mają rozkłady dyskretne, to przez ten symbol będziemy oznaczać łączną „funkcję prawdopodobieństwa"). Wtedy funkcję $L : \theta \to \mathbb{R}$

$$L(\theta) = f_\theta(x_1, x_2, \ldots, x_n) \tag{1.52}$$

nazywamy **wiarygodnością**.

Wiarygodność jest to właściwie to samo, co łączna gęstość $f_\theta(.)$. Inna nazwa ma podkreślać, iż interesujemy się tym razem *funkcją pewnego parametru* θ, przy ustalonych (bo przez nas już *zaobserwowanych*) wartościach próby.

Definicja 1.31. *Powiemy, że* $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ *jest* **estymatorem największej wiarygodności** *parametru* θ*, jeśli funkcja wiarygodności przyjmuje swoje maksimum w punkcie* $\hat{\theta}$*, tzn.*

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta) \ . \tag{1.53}$$

Symbolicznie to, że $\hat{\theta}$ jest estymatorem największej wiarygodności parametru θ, zapisujemy $\hat{\theta} = \mathrm{ENW}(\theta)$. Ponadto z definicji przyjmujemy, że jeśli rozważamy estymator $g(\theta)$ dla jakiejś znanej nam funkcji $g(.)$, to bezpośrednio $\mathrm{ENW}(g(\theta)) = g(\hat{\theta})$.

1.3 Twierdzenia graniczne

Twierdzenie 1.32 (Prawo wielkich liczb Kołmogorowa)**.** *Niech* $X_1, X_2, \ldots$ *będzie ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie (w skrócie* iid *– independent, identically distributed). Zbieżność*

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{p.n.} \mathbb{E}\, X \tag{1.54}$$

zachodzi wtedy i tylko wtedy, gdy istnieje $\mathbb{E}\, X$ *(wartość oczekiwana pojedynczej zmiennej* X_i*).*

Twierdzenie to jest istotne przy wykazaniu prawdziwości metod Monte Carlo.

Twierdzenie 1.33 (Centralne twierdzenie graniczne)**.** *Niech* $X_1, X_2, \ldots$ *będzie ciągiem zmiennych* iid *o wartości oczekiwanej pojedynczej zmiennej równej* $\mathbb{E}\, X$ *i skończonym odchyleniu standardowym* $\mathrm{D}\, X$*. Wtedy zachodzi zbieżność*

$$\frac{X_1 + X_2 + \ldots + X_n - n\,\mathbb{E}\, X}{\mathrm{D}\, X \sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1) \ . \tag{1.55}$$

Twierdzenie to pozwala oszacować błąd w metodach Monte Carlo.

1.4 Wybrane rozkłady prawdopodobieństwa

1.4.1 Rozkład dwupunktowy

Doświadczenie losowe ma tylko dwa możliwe wyniki, zazwyczaj zapisywane jako „1" i „0" (tak / nie, sukces / porażka, prawidłowy, nieprawidłowy, itd.). Prawdopodobieństwo „sukcesu" jest równe p, gdzie oczywiście $0 \leqslant p \leqslant 1$. Stąd rozkład prawdopodobieństwa dany jest wzorem

$$\mathrm{P}(X = 1) = p \ , \ \mathrm{P}(X = 0) = 1 - p \ . \tag{1.56}$$

Ważniejsze charakterystyki: $\mathbb{E}\, X = p, \mathrm{Var}\, X = p(1-p)$.

Zastosowanie: rzut monetą, kontrola sprawności pojedynczego elementu na linii produkcyjnej, zaliczenie egzaminu.

1.4.2 Rozkład dwumianowy

Załóżmy, że mamy n niezależnych powtórzeń takiego doświadczenia losowego, które ma tylko dwa możliwe wyniki (zwane tradycyjnie **porażką** i **sukcesem**). Oznacza to, że n razy powtarzamy doświadczenie z rozkładu dwupunktowego. Przez p, jak poprzednio, oznaczmy prawdopodobieństwo sukcesu w pojedynczej próbie. Wtedy prawdopodobieństwo zajścia k sukcesów w n próbach (czyli zdarzenia $X = k$) określone jest wzorem

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \ . \tag{1.57}$$

Ważniejsze charakterystyki: $\mathbb{E}\, X = np, \mathrm{Var}\, X = np(1-p)$. Tradycyjnie rozkład ten zapisujemy skrótowo $\mathrm{Bin}(n; p)$.

Zastosowanie: wielokrotne rzut monetą, kontrola sprawności n elementów na linii produkcyjnej, strzelanie do tarczy (trafienie / pudło).

1.4.3 Rozkład geometryczny

Załóżmy, że wykonujemy niezależne powtórzenia doświadczenia losowego, które ma tylko dwa możliwe wyniki, **aż do osiągnięcia sukcesu**. Przez p oznaczymy prawdopodobieństwo zajścia sukcesu w pojedynczej próbie. Wtedy liczba wykonanych doświadczeń ma rozkład geometryczny. Niech X będzie tą liczbą prób do momentu zajścia pierwszego sukcesu. Prawdopodobieństwo zdarzenia $X = k$ (czyli na początku nastąpiło $k-1$ porażek, a potem pierwszy sukces) dane jest wzorem

$$\mathrm{P}(X = k) = p(1-p)^{k-1} \ , \tag{1.58}$$

gdzie $k = 0, 1, \ldots$ Ważniejsze charakterystyki: $\mathbb{E}\, X = 1/p, \mathrm{Var}\, X = (1-p)/p^2$.

Zastosowanie: liczba rzutów monetą do momentu wypadnięcia pierwszego orła, liczba elementów na taśmie produkcyjnej zanim nie natrafimy na wadliwy, liczba wypełnionych losów TotoLotka, zanim po raz pierwszy nie trafimy „szóstki".

1.4.4 Rozkład Poissona

Jeśli zmienna pochodzi z rozkładu Poissona, to jej rozkład prawdopodobieństwa opisany jest wzorem

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1.59}$$

dla $k = 0, 1, \ldots$, gdzie $\lambda > 0$ jest parametrem tego rozkładu. Tradycyjnie rozkład ten oznaczamy skrótem Poiss(λ). Ważniejsze charakterystyki: $\mathbb{E}\, X = \lambda$, $\mathrm{Var}\, X = \lambda$.

Zastosowanie: ilość wypadków, do których doszło w pewnym ustalonym przedziale czasowym, liczba cząstek wyemitowanych przez radioaktywny materiał w pewnym przedziale czasowym, liczba zgłoszeń klientów w sieci w pewnym okresie (np. w ciągu godziny). Rozkład ściśle związany z rozkładem wykładniczym.

Jeśli dla ustalonego λ stworzymy wykres funkcji $\mathrm{P}(X = k)$ względem k (czyli wykres funkcji prawdopodobieństwa), będzie on malejący (tzn. wraz ze wzrostem k odpowiednie słupki pokazujące prawdopodobieństwo będą coraz niższe).

1.4.5 Rozkład jednostajny (równomierny)

Najprostszy z ciągłych rozkładów prawdopodobieństwa, oznaczany zazwyczaj skrotem $U\,[a; b]$. Jego gęstość na przedziale $[a; b]$ opisana jest wzorem

$$f(t) = \frac{1}{b - a}\ . \tag{1.60}$$

Oznacza to zatem, że prawdopodobieństwo zaobserwowania wartości zmiennej z dowolnego, małego przedziału o długości dx jest stałe i takie samo na całym przedziale $[a; b]$. Ważniejsze charakterystyki: $\mathbb{E}\, X = \frac{a+b}{2}$, $\mathrm{Var}\, X = \frac{(b-a)^2}{12}$.

Zastosowanie: przy „równomierności" zdarzeń na przedziale losowym, np. przypadkowy wybór liczby z przedziału $[0; 100000]$.

Wykres funkcji gęstości dla tego rozkładu jest stały na przedziale $[a; b]$ i równy zero poza nim.

1.4.6 Rozkład wykładniczy

Zmienna losowa X pochodzi z rozkładu wykładniczego (co zapisujemy $X \sim \mathrm{Exp}(\lambda)$), jeśli gęstość $f(.)$ jest równa

$$f(t) = \lambda e^{-\lambda t} \tag{1.61}$$

dla $t \geqslant 0$ i $f(t) = 0$ dla $t < 0$. Ważniejsze charakterystyki: $\mathbb{E}\, X = \frac{1}{\lambda}$, $\mathrm{Var}\, X = \frac{1}{\lambda^2}$.

Zastosowanie: jeśli w pewnym ustalonym przedziale czasowym liczba wystąpień jakiegoś zdarzenia jest zmienną z rozkładu Poissona, to okres czasu pomiędzy kolejnymi zdarzeniami jest właśnie zmienną z rozkładu wykładniczego, np. czas do wyemitowania radioaktywnej cząstki, czas do kolejnego zgłoszenia klienta w sieci.

Wykres funkcji gęstości dla tego rozkładu maleje (wykładniczo), począwszy od wartości przyjmowanej dla $t = 0$.

1.4.7 Rozkład normalny

Jeden z najważniejszych w statystyce rozkładów. Zmienna losowa X pochodzi z rozkładu normalnego (co zapisujemy $X \sim N(\mu, \sigma^2)$), jeśli gęstość $f(.)$ jest równa

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) , \tag{1.62}$$

gdzie $\sigma > 0$. Parametr μ nazywamy wartością oczekiwaną (lub średnią), a σ^2 – wariancją.

Zastosowanie: bardzo różnorodne zastosowania, np. modelowanie wzrostu osób, ocen z egzaminu, cen akcji, ilości opadów, temperatury, itd. Ogólnie rzecz biorąc, rozkład ten stosuje się wtedy, gdy pewna zmienna losowa jest wynikiem sumarycznego działania wielu „małych" i „niezależnych" czynników.

Wykres funkcji gęstości dla tego rozkładu ma postać dzwonu o maksimum w $t = \mu$.

1.4.8 Rozkład t-Studenta (rozkład t)

Rozkład bardzo często wykorzystywany w wielu testach statystycznych. Zmienna losowa X pochodzi z rozkładu t-Studenta (w skrócie *rozkładu t*, co zapisujemy $X \sim t(n)$), jeśli gęstość $f(.)$ ma postać

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} , \tag{1.63}$$

gdzie parametr $n \in \mathbb{N}_+$ zwany jest stopniami swobody (lub liczbą śladów).

Zastosowanie: testy statystyczne.

Ważniejsze charakterystyki (dla $n > 2$, dla mniejszej liczby stopni swobody niektóre momenty nie istnieją): $\mathbb{E}\, X = 0, \operatorname{Var} X = \frac{n}{n-2}$.

Dla $n \to \infty$ wykres gęstości tego rozkładu coraz bardziej przypomina gęstość standardowego rozkładu normalnego.

1.4.9 Rozkład χ^2 (chi-kwadrat)

Rozkład bardzo często wykorzystywany w wielu testach statystycznych. Zmienna losowa X pochodzi z rozkładu chi-kwadrat, co zapisujemy $X \sim \chi^2(n)$, jeśli gęstość $f(.)$ ma postać

$$f(t) = \frac{t^{\frac{n}{2}-1} e^{-\frac{t}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \tag{1.64}$$

dla $t > 0$, przy czym parametr $n \in \mathbb{N}_+$ zwany jest ilością śladów. Wykres gęstości ma postać „wolno przesuwającej się górki".

Ważniejsze charakterystyki: $\mathbb{E}\, X = n, \operatorname{Var} X = 2n$.

1.4.10 Notacja macierzowa. Wielowymiarowy rozkład losowy

Notacja macierzowa jest bardzo przydatna, gdy rozpatrujemy wielowymiarowe zmienne losowe. Takie wielowymiarowe zmienne losowe będziemy przed-

stawiać w postaci wektorów-kolumn, np.

$$Z = \begin{pmatrix} Z^{(1)} \\ \vdots \\ Z^{(k)} \end{pmatrix} , \tag{1.65}$$

gdzie Z jest zmienną losową (wektorem) złożoną z k „pojedynczych" zmiennych $Z^{(1)}, \ldots, Z^{(k)}$. Dla takiej zmiennej losowej – wektora, mamy odpowiednio $\mathbb{E}Z = \left(\mathbb{E}Z^{(1)}, \ldots, \mathbb{E}Z^{(k)}\right)^T$, czyli jego wartość oczekiwana jest wektorem poszczególnych wartości oczekiwanych. Znak T oznacza *transpozycję* wektora.

Macierzą kowariancji zmiennej Z jest macierz

$$\operatorname{VAR} Z = \begin{pmatrix} \operatorname{Var} Z^{(1)} & \operatorname{Cov}(Z^{(1)}, Z^{(2)}) & \ldots & \operatorname{Cov}(Z^{(1)}, Z^{(k)}) \\ \operatorname{Cov}(Z^{(2)}, Z^{(1)}) & \operatorname{Var} Z^{(2)} & \ldots & \operatorname{Cov}(Z^{(2)}, Z^{(k)}) \\ \ldots & & & \\ \operatorname{Cov}(Z^{(k)}, Z^{(1)}) & \operatorname{Cov}(Z^{(k)}, Z^{(2)}) & \ldots & \operatorname{Var} Z^{(k)} \end{pmatrix} , \tag{1.66}$$

czyli macierz, która na głównej przekątnej ma wariancje dla poszczególnych zmiennych losowych $Z^{(1)}, Z^{(2)}, \ldots, Z^{(k)}$, zaś poza nią są wszystkie możliwe kowariancje par „składowych" zmiennych losowych. Zachodzi przy tym

$$\operatorname{VAR} Z = \mathbb{E}(Z - \mathbb{E}\, Z)(Z - \mathbb{E}\, Z)^T , \tag{1.67}$$

jest to więc „uogólnienie" wzoru na wariancję dla jednowymiarowej zmiennej losowej na większą liczbę wymiarów. Macierz kowariancji musi być *macierzą symetryczną* (z symetrii samej kowariancji).

Wielowymiarowy rozkład normalny Z posiada gęstość daną wzorem

$$f_Z(t) = (2\pi)^{-n/2} \det \mathbb{W}^{-1} \exp\left[-\frac{1}{2}(t-\mu)^T \mathbb{W}^{-1}(t-\mu)\right] , \tag{1.68}$$

gdzie $\mathbb{W}$ jest macierzą symetryczną, dodatnio określoną o wymiarach $k \times k$, zaś μ jest wektorem k-wymiarowym. Przy tak zdefiniowanej gęstości $\mathbb{E}\, Z = \mu$ oraz $\operatorname{VAR} Z = \mathbb{W}$. Zgodnie z ogólną konwencją będziemy zatem pisać $Z \sim N(\mu, \mathbb{W})$.

Zauważmy, że jeśli $Z \sim N(\mu, \sigma^2\mathbb{I})$, gdzie $\mathbb{I}$ jest macierzą jednostkową, to kowariancja pomiędzy poszczególnymi zmiennymi równa się zero. Dodatkowo, z pewnych twierdzeń, jeśli zmienna Z pochodzi z rozkładu normalnego, to w takim przypadku poszczególne składowe $Z^{(1)}, Z^{(2)}, \ldots, Z^{(k)}$ są *niezależne*. Należy jednak pamiętać, iż nie jest to prawdą, gdy Z pochodzi z innego rozkładu.

1.5 Procesy stochastyczne

Definicja 1.34. *Procesem stochastycznym nazywamy rodzinę zmiennych losowych $S_t)_{t\in T}$ indeksowanych parametrem t, określonych na tej samej przestrzeni probabilistycznej. Zbiór $\mathcal{S}$, t.ż dla każdego $t \in T$ wartości S_t będą zawierać się w $\mathcal{S}$, nazywamy przestrzenią stanów procesu stochastycznego. Wartość pojedynczej zmiennej losowej S_t nazywamy stanem procesu stochastycznego.*

Zazwyczaj $T = [0; +\infty)$ i indeks t bywa utożsamiany z *czasem*, a sam proces stochastyczny – z *ewolucją pewnego zjawiska w czasie*. Procesy stochastyczne zdefiniowane na dyskretnej przestrzeni stanów nazywane są czasami *łańcuchami*.

Przyjrzymy się przykładom kilku ważnych procesów stochastycznych.

Definicja 1.35. *Proces Poissona $\mathcal{N}(\lambda)$ o parametrze $\lambda > 0$ to proces stochastyczny dla $T = [0; +\infty)$, który spełnia warunki:*

1. *$\mathcal{N}_0 = 0$ z prawdopodobieństwem równym jeden,*
2. *$\mathcal{N}_t$ ma przyrosty niezależne i nieujemne (tzn. dla dowolnych $0 \leqslant s < t \leqslant u < w$ zmienne losowe $\mathcal{N}_t - \mathcal{N}_s$ i $\mathcal{N}_w - \mathcal{N}_u$ są niezależne i nieujemne),*
3. *dla dowolnych $0 \leqslant s \leqslant t$ zachodzi $\mathcal{N}_t - \mathcal{N}_s \sim \text{Poiss}(\lambda(t-s))$,*

Proces ten nie ma pamięci, tzn. proces obserwowany od ustalonego miejsca jest nową kopią procesu Poissona i wcześniej zaszłe zdarzenia nie wpływają na prawdopodobieństwo zajścia zdarzenia w następnym okresie.

Definicja 1.36. *Proces stochastyczny $\{\mathcal{W}_t\}_{t \geqslant 0}$ nazywamy* **standardowym procesem Wienera**, *gdy spełnia następujące warunki:*

1. *$\mathcal{W}_0 = 0$ z prawdopodobieństwem równym jeden,*
2. *$\mathcal{W}_t$ ma przyrosty niezależne (tzn. dla dowolnych $0 \leqslant s < t \leqslant u < w$ zmienne losowe $\mathcal{W}_t - \mathcal{W}_s$ i $\mathcal{W}_w - \mathcal{W}_u$ są niezależne),*
3. *dla dowolnych $0 \leqslant s \leqslant t$ zachodzi $\mathcal{W}_t - \mathcal{W}_s \sim N(0, t-s)$,*
4. *trajektorie procesu $\mathcal{W}_t$ są ciągłe prawie na pewno (tzn. z prawdopodobieństwem jeden).*

Proces Wienera zwany jest również **ruchem Browna** i bywa rozpatrywany w zastosowaniach fizycznych oraz finansowych (patrz rozdział 5.3).

Procesami stochastycznymi są również łańcuchy Markowa (patrz rozdział 1.6).

1.6 Łańcuchy Markowa

1.6.1 Dyskretne łańcuchy Markowa

Niech $(X_i)_{i=0} = (X_0 = x_0, X_1, \ldots, X_n, \ldots)$ będzie oznaczać ciąg zmiennych losowych, t.ż. dla każdego i wartości X_i zawierać się będą w pewnym przeliczalnym zbiorze $\mathcal{X}$, zwanym dalej *przestrzenią stanów*, gdzie $\mathcal{X} \subset \mathbb{N}$. Moc zbioru $\mathcal{X}$ oznaczać będziemy dalej przez s_X. Wartość x_0, czyli punkt startowy ciągu $(X_i)_{i=0}$, nazywać będziemy jego *stanem początkowym*. Jeśli stan początkowy nie został wybrany deterministycznie, ale $X_0 \sim \pi_{X_0}$ dla pewnego rozkładu π_{X_0}, to ów rozkład nazwiemy *rozkładem początkowym*.

Definicja 1.37. *Ciąg $(X_i)_{i=0}$ nazwiemy* łańcuchem Markowa, *jeśli dla każdego $k > 0$ i dowolnego $\mathcal{A} \subset \mathcal{X}$ zachodzi własność:*

$$\mathrm{P}\left(X_k \in \mathcal{A} | X_{k-1}, \ldots, X_0\right) = \mathrm{P}\left(X_k \in \mathcal{A} | X_{k-1}\right) \ . \tag{1.69}$$

Innymi słowy, łańcuch Markowa posiada swoistą własność „braku pamięci długotrwałej". Oznacza to, że dla wiedzy o „przyszłym" zachowaniu się ciągu w chwili k i następnych, istotne jest tylko posiadanie informacji o chwili „teraźniejszej", czyli $k-1$. Nie jest natomiast istotna wiedza o „przeszłości" ciągu, a więc krokach $0, 1, \ldots, k-2$.

Dla uproszczenia notacji zamiast słów „. . . jest łańcuchem Markowa", używać będziemy skrótu „. . . jest ŁM". Ponieważ przestrzeń stanów $\mathcal{X}$, dla której zdefiniowaliśmy ŁM w tym rozdziale, jest przestrzenią dyskretną (tzn. skończoną lub przeliczalną), to w skrócie będziemy mówić o „dyskretnych ŁM", przypominając w ten sposób o mocy przestrzeni stanów.

Przykład 1.38. *Zauważmy, że ŁM jest w szczególności dowolny ciąg zmiennych* iid *(tzn. niezależnych zmiennych losowych, z których każda posiada ten sam rozkład prawdopodobieństwa). Bardziej interesującym jest prosty przykład dwustanowego ŁM o stanach oznaczonych przez* $\{1, 2\}$ *i prawdopodobieństwach przejściach pomiędzy stanami danymi przez*

$$\mathrm{P}(X_k = 1|X_{k-1} = 1) = p_{11} \ , \ \mathrm{P}(X_k = 2|X_{k-1} = 1) = p_{12} = 1 - p_{11} \ , \quad (1.70)$$
$$\mathrm{P}(X_k = 1|X_{k-1} = 2) = p_{21} \ , \ \mathrm{P}(X_k = 2|X_{k-1} = 2) = p_{22} = 1 - p_{21} \ , \quad (1.71)$$

gdzie $p_{11}, p_{21} \in [0; 1]$.

Definicja 1.39. *Powiemy, że łańcuch Markowa* $(X_i)_{i=0}$ *jest* jednorodny *(w j. ang.* homogeneous*), jeśli istnieje macierz* $\mathbb{P}_X$*, t.ż.*

$$\mathbb{P}_X = \left(\mathrm{P}\left(X_{k+1} = j|X_k = i\right)\right)_{i,j=1}^{s_X} \ , \quad (1.72)$$

niezależna od wartości kroku k*. Macierz taką będziemy nazywać* jednorodną macierzą przejścia *dla łańcucha* $(X_i)_{i=0}$.

Jednorodny ŁM charakteryzuje się bardzo istotną własnością – prawdopodobieństwa przejścia pomiędzy stanami nie zmieniają się w zależności od upływającego czasu, tzn. w każdym kroku pozostają takie same i są jednoznacznie określone przez macierz $\mathbb{P}_X$. Od tej chwili wszystkie ŁM, którymi zajmować się będziemy w tej pracy, będą jednorodnymi ŁM. W związku z tym, ilekroć będziemy też wspominali o macierzy przejścia dla ŁM, będziemy mieć na myśli jednorodną macierz przejścia opisaną w definicji 1.39.

Przykład 1.40. *W szczególności ŁM opisany prawdopodobieństwami przejścia* (1.70) – (1.71) *jest jednorodnym ŁM o macierzy przejścia*

$$\mathbb{P}_X = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \ . \quad (1.73)$$

Definicja 1.41. *Powiemy, że łańcuch Markowa posiada* rozkład stacjonarny π_X*, jeśli istnieje rozkład prawdopodobieństwa* π_X *dla którego zachodzi własność*

$$\pi_X^T = \pi_X^T \mathbb{P}_X \ , \quad (1.74)$$

gdzie zapis π_X *oznacza wektor – kolumnę.*

Rozkład stacjonarny pełni dla ŁM bardzo istotną rolę. Jest to mianowicie, przy spełnieniu pewnych dalej omówionych założeń (patrz rozdział 1.6.3), rozkład graniczny dla ŁM, niezależny od rozkładu lub stanu początkowego ciągu. Oznacza to, że jeśli owe założenia zostaną spełnione, możemy znaleźć rozkład graniczny dla danego ŁM wyznaczając jego rozkład stacjonarny ze wzoru (1.74).

Definicja 1.42. *Powiemy, że stan $b \in \mathcal{X}$ jest* osiągalny *ze stanu $a \in \mathcal{X}$, jeśli istnieje $n \in \mathbb{N}$, t.ż.* $\mathrm{P}(X_n = b|X_0 = a) > 0$. *Powiemy, że stany a i b* komunikują się*, jeśli stan a jest osiągalny z b i stan b jest osiągalny z a. Łańcuch Markowa nazwiemy* nieprzywiedlnym *(w j. ang.* irreducible*), gdy wszystkie stany przestrzeni stanów $\mathcal{X}$ komunikują się wzajemnie.*

Przykład 1.43. *Jako przykład rozpatrzmy trzystanowy ŁM $\{1, 2, 3\}$ opisany następującą macierzą przejścia*

$$\mathbb{P}_X = \begin{pmatrix} 0,2 & 0,3 & 0,5 \\ 0,4 & 0,3 & 0,3 \\ 0 & 0 & 1 \end{pmatrix} \ . \tag{1.75}$$

Jak łatwo zauważyć, np. stany $\{1, 2\}$ komunikują się wzajemnie. Stan $\{3\}$ jest osiągalny ze stanów $\{1, 2\}$, ale stany $\{1, 2\}$ nie są osiągalne ze stanu $\{3\}$, dlatego ŁM o macierzy (1.75) *nie jest nieprzywiedlny.*

Jeśli ŁM jest nieprzywiedlny, oznacza to w szczególności, że z dodatnim prawdopodobieństwem możemy przejść pomiędzy dowolnymi dwoma stanami po pewnej liczbie kroków zależnej od tych dwóch stanów.

Definicja 1.44. *Niech $a \in \mathcal{X}$. Stan a nazwiemy* powracającym *(w j. ang.* recurrent*), jeśli*

$$\mathrm{P}\left(\bigcup_{k=1}^{\infty} \{X_k = a\} \middle| X_0 = a \right) = 1 \ , \tag{1.76}$$

a chwilowym *(w j. ang.* transient*), gdy*

$$\mathrm{P}\left(\bigcup_{k=1}^{\infty} \{X_k = a\} \middle| X_0 = a \right) < 1 \ . \tag{1.77}$$

To, że stan a jest powracający, intuicyjnie oznacza, iż ŁM powraca do niego nieskończenie wiele razy. Jeśli stan jest chwilowy, to ŁM po pewnym, skończonym czasie opuści ten stan, aby więcej już do niego nie powrócić.

Def. 1.44 możemy także wyrazić w inny sposób, wprowadzając dodatkową funkcję zliczającą liczbę wejść ŁM do wybranego stanu $a \in \mathcal{X}$:

$$\eta_a = \sum_{k=0}^{\infty} \mathbb{1}(X_k = a) \ , \tag{1.78}$$

gdzie

$$\mathbb{1}(\text{warunek}) = \begin{cases} 0 & \text{gdy warunek jest fałszywy} \\ 1 & \text{gdy warunek jest prawdziwy} \end{cases} \tag{1.79}$$

jest tzw. funkcją charakterystyczną. Warunki równoważne def. 1.44 podają wtedy dwa następujące twierdzenia.

Twierdzenie 1.45. *Stan $a \in \mathcal{X}$ jest powracający wtedy i tylko wtedy, gdy*

$$\mathrm{P}\left(\eta_a = \infty | X_0 = a\right) = 1 \ . \tag{1.80}$$

Stan $a \in \mathcal{X}$ jest chwilowy wtedy i tylko wtedy, gdy

$$\mathrm{P}\left(\eta_a < \infty | X_0 = a\right) = 1 \ . \tag{1.81}$$

Twierdzenie 1.46. *Niech $a \in \mathcal{X}$. Stan $a \in \mathcal{X}$ jest powracający wtedy i tylko wtedy, gdy*

$$\mathbb{E}\left(\eta_a | X_0 = a\right) = \infty \ . \tag{1.82}$$

Stan $a \in \mathcal{X}$ jest chwilowy wtedy i tylko wtedy, gdy

$$\mathbb{E}\left(\eta_a | X_0 = a\right) < \infty \ . \tag{1.83}$$

Dowody powyższych twierdzeń znaleźć można np. w [19].

Jak łatwo zauważyć, jeśli przestrzeń stanów $\mathcal{X}$ jest skończona, a ŁM – nieprzywiedlny, to wszystkie stany tego ŁM muszą być powracające.

Twierdzenie 1.47. *W nieprzywiedlnym ŁM wszystkie stany są tego samego typu – jeśli jeden jest powracający, to wszystkie są powracające. Jeśli jeden jest chwilowy, to wszystkie są chwilowe.*

Dzięki powyższemu twierdzeniu możemy jednoznacznie mówić o „powracających" i „chwilowych" ŁM w najbardziej interesującym nas przypadku nieprzywiedlnych ŁM. Dowód powyższego twierdzenia znaleźć można np. w [19].

Definicja 1.48. Okresem *stanu $a \in \mathcal{X}$ nazwiemy liczbę*

$$o(a) = \text{NWD}\ \{n \in \mathbb{N} : \mathrm{P}(X_n = a | X_0 = a) > 0\} \ . \tag{1.84}$$

Stan a nazwiemy okresowym *(w j. ang.* periodical*) o okresie $o(a)$, gdy $o(a) > 1$ i* nieokresowym *(w j. ang.* aperiodical*), gdy $o(a) = 1$.*

Przykład 1.49. *Przykładowo, „skaczący" dwustanowy ŁM zdefiniowany macierzą przejścia*

$$\mathbb{P}_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{1.85}$$

jest ŁM o okresie dwa – łańcuch w pierwszym kroku każdego cyklu przeskakuje do innego stanu, aby potem, w drugim kroku cyklu, wrócić do stanu w którym znajdował się na początku. Jak łatwo zauważyć, dla dowolnego stanu tego łańcucha liczba kroków o niezerowym prawdopodobieństwie zdefiniowanym przez (1.84) *wynosi odpowiednio 2,4,6, itd.*

Twierdzenie 1.50. *W nieprzywiedlnym łańcuchu Markowa wszystkie stany mają ten sam okres lub łańcuch jest nieokresowy.*

Dowód powyższego twierdzenia znaleźć można np. w [19].

Dzięki tw. 1.50 możemy pojęcie „okresowości" ŁM sprowadzić zawsze do zagadnienia dotyczącego wszystkich stanów tego ŁM jednocześnie. Umożliwia to mówienie o „nieokresowym ŁM", czyli ŁM o wszystkich stanach nieokresowych.

W literaturze dotyczącej ŁM wykorzystywana jest często następująca, bardzo przydatna notacja. Wprowadźmy mianowicie symbole

$$\mathrm{P}_{x_0}(.) = \mathrm{P}(.|X_0 = x_0) \tag{1.86}$$

i

$$\mathrm{P}^n(y|X_0 = x_0) = \mathrm{P}(X_n = y|X_0 = x_0) \ , \tag{1.87}$$

gdzie, jak widzimy, $\mathrm{P}_{x_0}(.)$ jest prawdopodobieństwem wystąpienia pewnego zdarzenia pod warunkiem startu ze zdeterminowanego stanu początkowego x_0, a

$\mathrm{P}^n(y|X_0 = x_0)$ jest prawdopodobieństwem osiągnięcia stanu y w dokładnie n krokach, o ile startujemy ze stanu x_0.

W dalszej części pracy zakładać będziemy, że jeśli omawiamy dowolny dyskretny ŁM, to jest on jednocześnie jednorodny, nieprzywiedlny, powracający i nieokresowy w myśl powyższych definicji, oraz posiada określoną macierz przejścia i związany z nią rozkład stacjonarny.

Podejście takie wynika z faktu, że w metodach MCMC interesujemy się ŁM o właśnie takich własnościach, które wynikają zazwyczaj z samej konstrukcji algorytmu MCMC (patrz rozdział 6).

1.6.2 Łańcuchy Markowa o wartościach w przestrzeni ciągłej

W niniejszym rozdziale przeniesiemy definicje z rozdziału 1.6.1 na grunt łańcuchów Markowa o wartościach w przestrzeni ciągłej (np. na prostej rzeczywistej $\mathbb{R}$).

Niech $(X_i)_{i=0} = (X_0 = x_0, X_1, \ldots, X_n, \ldots)$ będzie oznaczać ciąg zmiennych losowych, t.ż. dla każdego i wartości X_i zawierać się będą w pewnym nieprzeliczalnym zbiorze $\mathcal{X}$, zwanym przestrzenią stanów, gdzie $\mathcal{X} \subset \mathbb{R}^p$ dla pewnego ustalonego $p \in \mathbb{N}$. Tak jak poprzednio, moc przestrzeni $\mathcal{X}$ oznaczać będziemy przez s_X.

Niech $\mathbb{B}(\mathcal{X})$ oznaczać będzie σ-ciało zbiorów borelowskich dla przestrzeni $\mathcal{X}$ (patrz np. [3]).

Definicja 1.51. *Ciąg $(X_i)_{i=0}$ nazwiemy* łańcuchem Markowa*, jeśli dla każdego $k > 0$ i dowolnego $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ zachodzi własność:*

$$\mathrm{P}\left(X_k \in \mathcal{A} | X_{k-1}, \ldots, X_0\right) = \mathrm{P}\left(X_k \in \mathcal{A} | X_{k-1}\right) \ . \tag{1.88}$$

Definicja 1.52. *Funkcję $\mathcal{K}_X : \mathcal{X} \times \mathbb{B}(\mathcal{X}) \times \mathbb{N} \to \mathbb{R}$ nazwiemy* jądrem przejścia *(w j. ang.* kernel*) dla ŁM $(X_i)_{i=0}$, jeśli:*

1. *dla każdego $x \in \mathcal{X}$, $\mathcal{K}_X(x, ., .)$ jest gęstością prawdopodobieństwa,*
2. *dla dowolnego $\mathcal{A} \in \mathbb{B}(\mathcal{X})$, $\mathcal{K}_X(., \mathcal{A}, .)$ jest funkcją mierzalną.*

Jądro przejścia $\mathcal{K}_X(x, y, k)$ możemy intuicyjnie utożsamiać z rodziną gęstości prawdopodobieństwa, zadających prawdopodobieństwo przejścia ze stanu x do stanu y w kroku k.

Dzięki temu możemy zdefiniować związek pomiędzy jądrem przejścia $\mathcal{K}_X(., ., .)$ a prawdopodobieństwem przejścia w kroku k-tym dla ŁM $(X_i)_{i=0}$ jako dany zależnością

$$\mathrm{P}(X_{k+1} \in \mathcal{B} | X_k = x) = \int_{\mathcal{B}} \mathcal{K}_X(x, y, k)\, dy \tag{1.89}$$

dla dowolnego $\mathcal{B} \in \mathbb{B}(\mathcal{X})$.

Jednym z prostszych przykładów ŁM o wartościach w przestrzeni ciągłej jest ruch Browna z czasem dyskretnym. Niech dla dowolnych $x, y \in \mathbb{R}$ i $k \in \mathbb{N}$

$$\mathcal{K}_X(x, y, k) = \varphi_{N(x;1)}(y) \ , \tag{1.90}$$

gdzie $\varphi_{N(x;1)}(y)$ jest gęstością w punkcie y rozkładu normalnego o średniej x i wariancji 1. Wtedy, z (1.89), mamy dla tego przykładu

$$\mathrm{P}(X_{k+1} \in \mathcal{B} | X_k = x) = \phi_{N(x;1)}(\mathcal{B}) \ , \tag{1.91}$$

gdzie $\phi_{N(x;1)}(\mathcal{B})$ jest prawdopodobieństwem zajścia zdarzenia $\mathcal{B}$ dla rozkładu normalnego o średniej x i wariancji 1.

Następne definicje są przeniesieniem poszczególnych własności dyskretnych ŁM na ŁM o ciągłej przestrzeni stanów.

Definicja 1.53. *Łańcuch Markowa* $(X_i)_{i=0}$ *nazwiemy* jednorodnym*, jeśli postać jądra przejścia* $\mathcal{K}_X(.,.,.)$ *dla tego łańcucha nie zależy od wartości kroku* k*.*

Jak łatwo zauważyć, ŁM zdefiniowany jądrem przejścia (1.90) jest jednorodnym ŁM.

Od tej chwili będziemy rozpatrywali tylko łańcuchy jednorodne. Dla ułatwienia notacji jądro przejścia zapisywać będziemy od tej pory jako funkcję $\mathcal{K}_X(.,.)$ dwu zmiennych postaci $\mathcal{K}_X : \mathcal{X} \times \mathbb{B}(\mathcal{X}) \to \mathbb{R}$, tzn. zależną jedynie od wartości stanów, a nie numeru kroku.

Definicja 1.54. *Łańcuch Markowa* $(X_i)_{i=0}$ *nazwiemy* ρ–nieprzywiedlnym*, jeśli istnieje miara* ρ *na przestrzeni* $\mathcal{X}$*, t.ż. dla każdego* $\mathcal{A} \in \mathbb{B}(\mathcal{X})$*, dla którego* $\rho(\mathcal{A}) > 0$*, istnieje* n*, t.ż.*

$$\mathrm{P}(X_n \in \mathcal{A} | X_0) > 0 \ . \tag{1.92}$$

Powyższa definicja jest w swej istocie niemal identyczna z definicją nieprzywiedlności dla dyskretnych ŁM, gdzie dla dowolnych dwóch stanów istnieje niezerowe prawdopodobieństwo przejścia pomiędzy tymi stanami w pewnej liczbie kroków. Jednak, z oczywistych powodów, dla ŁM o ciągłej przestrzeni stanów prawdopodobieństwo znalezienia się w pojedynczym, wybranym stanie jest zazwyczaj równe zero. Stąd potrzeba wprowadzenia w definicji 1.54 dodatkowej miary ρ, względem której definiujemy prawdopodobieństwa zdarzeń $\mathcal{A} \in \mathbb{B}(\mathcal{X})$. Co istotne, własność nieprzywiedlności dla ŁM nie zależy od wyboru miary ρ, jest więc cechą samego ŁM (patrz [30], rozdział 6.3.1). Stąd możemy także zupełnie jednoznacznie mówić o „nieprzywiedlnych ŁM" określonych na ciągłej przestrzeni stanów. Należy jednak pamiętać, że w następnych twierdzeniach i definicjach nieodzowne jest *istnienie* pewnej miary ρ, względem której określamy stany o niezerowym prawdopodobieństwie.

Na przykład dla ŁM opisanego jądrem przejścia (1.90) istnieje niezerowe, choć być może bardzo małe, prawdopodobieństwo przejścia nawet w pojedynczym kroku. Tak więc jest to łańcuch nieprzywiedlny.

Definicja 1.55. *Rozkład prawdopodobieństwa* π_X *nazwiemy* rozkładem stacjonarnym *dla* $(X_i)_{i=0}$ *będącego ŁM, jeśli dla każdego* $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ *spełnia on warunek:*

$$\pi_X(\mathcal{A}) = \int_{\mathcal{A}} \int_{\mathcal{X}} \mathcal{K}_X(x, y) \, d\pi_X(x) \, dy \ . \tag{1.93}$$

Jak łatwo zauważyć, jądro przejścia $\mathcal{K}_X(.,.)$ spełnia podobną rolę, co macierz przejścia $\mathbb{P}_X$ dla rozkładu stacjonarnego w przypadku dyskretnych ŁM (patrz definicja 1.41).

Definicja 1.56. *Zbiór* $\mathcal{C} \subset \mathcal{X}$ *nazwiemy* zbiorem małym *(w j. ang.* small set*), jeśli istnieje* $m \in \mathbb{N}$ *i miara* $\nu_m > 0$*, t.ż. dla każdego* $x \in \mathcal{C}$ *i każdego* $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ *zachodzi*

$$\mathrm{P}(X_m \in \mathcal{A} | X_0 = x) \geqslant \nu_m(\mathcal{A}) \ . \tag{1.94}$$

Definicję powyższą wraz z opisem szeregu jej własności znaleźć można np. w [31].

Zbiór mały $\mathcal{C}$ ma, w myśl definicji 1.56, specyficzną własność. Istnieje pewna ustalona liczba kroków m oraz oddzielona od zera miara ν_m, taka, że prawdopodobieństwo przejścia pomiędzy dowolnym stanem $x \in \mathcal{C}$ i dowolnym zdarzeniem $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ jest ograniczone od dołu właśnie przez miarę ν_m. Innymi słowy, dowolne zdarzenie $\mathcal{A}$ jest osiągalne po pewnej liczbie kroków ze zbioru małego $\mathcal{C}$. Należy przy tym zauważyć, że liczba tych kroków m *nie jest zależna* od zbioru $\mathcal{A}$. Wbrew pozorom, zbiory małe występują w ŁM bardzo powszechnie. Jeśli tylko ŁM jest dostatecznie regularny (tzn. jest nieprzywiedlny), to można jego przestrzeń stanów $\mathcal{X}$ zdekomponować na sumę zbiorów małych (patrz [27]).

Definicja 1.57. *Powiemy, że nieprzywiedlny ŁM* $(X_i)_{i=0}$ *ma* okres długości d, *jeśli istnieje zbiór mały* $\mathcal{C}$, $M \in \mathbb{N}$ *i rozkład prawdopodobieństwa* ν_M, *t.ż.* d *jest NWD dla*

$$\{m \geqslant 1 : \exists\ \delta_m > 0 \text{ } t.ż.\ \mathcal{C} \text{ jest mały dla pewnego rozkładu prawdopodobieństwa } \nu_m \geqslant \delta_m \nu_M\}\ . \quad (1.95)$$

ŁM nazwiemy nieokresowym, jeśli $d = 1$.

Mimo skomplikowanej postaci, definicja powyższa jest właściwie analogiczna do definicji 1.48 – okresowości dla dyskretnych ŁM. Zbiór mały $\mathcal{C}$ pełni rolę stanu, dla którego sprawdzamy NWD długości cykli powrotu do tego zbioru. Co istotne, okresowość ŁM nie zależy od wyboru zbioru małego $\mathcal{C}$. Oznacza to, że okresowość (lub nieokresowość) jest cechą charakterystyczną całego ŁM (patrz [30], rozdział 6.3.3), dzięki czemu możemy mówić o okresowych (lub nieokresowych) ŁM w sposób zupełnie jednoznaczny. Jak więc widzimy, sytuacja jest zupełnie analogiczna do przypadku jednoznacznej definicji okresowości dla dyskretnych ŁM.

Niech $\mathcal{A} \in \mathbb{B}(\mathcal{X})$. Wtedy, analogicznie do (1.78), przez $\eta_{\mathcal{A}}$ oznaczać będziemy liczbę wejść łańcucha $(X_i)_{i=0}$ do zbioru $\mathcal{A}$, tzn.

$$\eta_{\mathcal{A}} = \sum_{k=0}^{\infty} \mathbb{1}(X_k \in \mathcal{A})\ . \quad (1.96)$$

W przypadku ŁM określonych na przestrzeni ciągłej, kwestia *powracalności stanu* jest bardziej skomplikowana niż dla dyskretnych ŁM. Przyjrzymy się teraz bliżej temu zagadnieniu.

Definicja 1.58. *Zbiór* $\mathcal{A}$ *nazwiemy* powracającym, *jeśli dla każdego* $x \in \mathcal{A}$ *zachodzi*

$$\mathbb{E}_x \eta_{\mathcal{A}} = \infty\ . \quad (1.97)$$

Łańcuch Markowa $(X_i)_{i=0}$ *nazwiemy* powracającym, *jeśli istnieje miara* ρ, *t.ż. łańcuch jest nieprzywiedlny, i dla każdego zbioru* $\mathcal{A}$, *t.ż.* $\rho(\mathcal{A}) > 0$, *zbiór* $\mathcal{A}$ *jest powracający.*

Definicja 1.59. *Zbiór* $\mathcal{A}$ *nazwiemy* powracającym w sensie Harrisa *(w j. ang.* Harris recurrent*), jeśli dla każdego* $x \in \mathcal{A}$ *zachodzi*

$$\mathrm{P}(\eta_{\mathcal{A}} = \infty | X_0 = x) = 1\ . \quad (1.98)$$

Łańcuch Markowa $(X_i)_{i=0}$ *nazwiemy* powracającym w sensie Harrisa*, jeśli istnieje miara* ρ*, t.ż. łańcuch jest nieprzywiedlny, i dla każdego zbioru* $\mathcal{A}$*, t.ż.* $\rho(\mathcal{A}) > 0$*, zbiór* $\mathcal{A}$ *jest powracający w sensie Harrisa.*

Zauważmy pewną dychotomię pomiędzy powyższymi definicjami. O ile w przypadku dyskretnych ŁM warunek pewnego powrotu (1.80) był równoważny nieskończoności wartości oczekiwanej liczby powrotów (1.82) dzięki twierdzeniom 1.45 i 1.46, to w przypadku ŁM zdefiniowanych na przestrzeni ciągłej powracalność w sensie definicji 1.58 nie jest równoważna powracalności w sensie Harrisa w twierdzeniu 1.59. Wynika to z poniższych twierdzeń.

Twierdzenie 1.60. *Jeśli istnieje miara* ρ*, t.ż. łańcuch jest nieprzywiedlny i zbiór mały* $\mathcal{C}$ *dla którego* $\rho(\mathcal{C}) > 0$ *i dla dowolnego* $x \in \mathcal{C}$ *spełniony jest warunek*

$$\mathrm{P}_x\left(\inf_k\{X_k \in \mathcal{C}\} < \infty\right) = 1 \ , \tag{1.99}$$

to łańcuch Markowa $(X_i)_{i=0}$ *jest powracający,*

Twierdzenie 1.61. *Jeśli dla każdego* $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ *i dowolnego* $x \in \mathcal{A}$ *zachodzi*

$$\mathrm{P}_x\left(\inf_k\{X_k \in \mathcal{A}\} < \infty\right) = 1 \ , \tag{1.100}$$

to ŁM jest powracający w sensie Harrisa.

Dowody powyższych twierdzeń znaleźć można w [30]. Jak widzimy, dla powracających ŁM wymagamy jedynie istnienia pewnego zbioru małego $\mathcal{C}$, dla którego spełniony jest warunek (1.99). Własność powracalności w sensie Harrisa jest silniejsza, gdyż podobny warunek (1.100) musi być spełniony dla *dowolnego* zbioru $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ i punktu $x \in \mathcal{A}$.

Własność powracalności w sensie Harrisa jest bardzo istotna dla ŁM określonych na przestrzeni ciągłej. Jeśli ŁM jest powracający w sensie Harrisa, to będzie odwiedzać każdy zbiór $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ nieskończenie wiele razy z prawdopodobieństwem jeden, dla dowolnie wybranego stanu początkowego x. Oczywiście, takie podejście jest bardzo ważne dla metod MCMC, gdzie cały ciąg startuje z pewnego *pojedynczego* stanu (patrz rozdział 6) i chcemy, aby „odwiedził" on całą przestrzeń stanów.

Podobnie do umowy z poprzedniego rozdziału, w dalszej części pracy zakładać będziemy, że jeśli omawiamy dowolny ŁM określony na ciągłej przestrzeni stanów, to jest on jednocześnie jednorodny, nieprzywiedlny, nieokresowy i powracający w sensie Harrisa w myśl powyższych definicji, oraz posiada określone jądro przejścia i związany z nim rozkład stacjonarny.

1.6.3 Własność Markowa i twierdzenia ergodyczne

W niniejszym rozdziale omówimy pojęcie własności Markowa, które w istotny sposób wykorzystywane będzie w dalszej części pracy. Ponadto przedstawimy pokrótce najważniejsze z tzw. twierdzeń ergodycznych, które opisują graniczne własności ŁM.

Definicja 1.62. *Niech* T *będzie pewnym określonym ciągiem, a* $(\Omega, \mathcal{F}, P)$ *– ustaloną przestrzenią probabilistyczną. Wtedy* filtracją *nazywamy niemalejącą rodzinę* σ*-ciał* $(\mathcal{F}_t)_{t\in T}$*, gdzie* $\mathcal{F}_t \subset \mathcal{F}$ *dla każdego* $t \in T$*.*

W szczególności dla ciągu zmiennych losowych $(X_i)_{i=0}$, rozważanego w tej pracy, możemy zdefiniować naturalną filtrację $(\mathcal{F}_i)_{i=0}$, gdzie $\mathcal{F}_i = \sigma(X_s : s \leqslant i)$. Innymi słowy, σ-ciało $\mathcal{F}_k$ zawiera całą wiedzę o własnościach ciągu $(X_i)_{i=0}^k$, czyli do kroku k włącznie. W oczywisty sposób zmienna losowa X_k jest wtedy $\mathcal{F}_k$--mierzalna dla każdego $k = 0, 1, \ldots$.

Definicja 1.63. Momentem stopu (momentem zatrzymania) *względem filtracji* $(\mathcal{F}_t)_{t\in T}$ *nazywamy zmienną losową* $\tau : \Omega \to T \cup \{+\infty\}$*, spełniającą warunek*

$$\{\tau \leqslant t\} \in \mathcal{F}_t \tag{1.101}$$

dla wszystkich $t \in T$*.*

Powyższa definicja intuicyjnie oznacza, że do podjęcia decyzji, czy moment stopu τ już wystąpił, czy też jeszcze nie zaszedł do momentu t, wystarczy nam zawsze posiadanie wiedzy o rodzinie σ-ciał tylko do chwili t włącznie. Przykładami momentów stopu dla ciągu $(X_i)_{i=0}$ są np. operatory minimum – $Z_k^{\min} = \min_{s\leqslant k}\{X_s\}$, i maksimum – $Z_k^{\max} = \max_{s\leqslant k}\{X_s\}$.

Twierdzenie 1.64. *Niech* $(X_i)_{i=0}$ *będzie ŁM o przestrzeni stanów* $\mathcal{X}$ *i niech* $\mathcal{A} \subset \mathbb{B}(\mathcal{X})$*. Wtedy zmienne losowe*

$$\zeta_1 := \min\{i = 1, \ldots : X_i \in \mathcal{A}\}\ , \tag{1.102}$$

$$\zeta_{k+1} := \min\{i > \zeta_k : X_i \in \mathcal{A}\} \tag{1.103}$$

są momentami stopu względem naturalnej filtracji dla tego ŁM.

Oznacza to, że momenty pierwszego wejścia i kolejnych powrotów do ustalonego zbioru $\mathcal{A}$ są momentami stopu. Dowód powyższego twierdzenia znaleźć można np. w [4], rozdział 2.7.

Twierdzenie 1.65 (Słaba własność Markowa)**.** *Niech* $(X_i)_{i=0}$ *będzie ŁM o przestrzeni stanów* $\mathcal{X}$*. Wtedy dla dowolnego ustalonego* n *i* $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ *zachodzi*

$$\begin{aligned}\mathrm{P}\left(\{(X_{n+1}, X_{n+2}, \ldots) \in \mathcal{A}\} | X_n = x, \ldots, X_0\right) = \\ = \mathrm{P}\left(\{(X_{n+1}, X_{n+2}, \ldots) \in \mathcal{A}\} | X_n = x\right) = \\ = \mathrm{P}\left(\{(X_1, X_2, \ldots) \in \mathcal{A}\} | X_0 = x\right)\ .\end{aligned} \tag{1.104}$$

Jak łatwo zauważyć, słaba własność Markowa jest uogólnieniem własności ŁM, którą widzieliśmy w definicjach 1.37 i 1.51. Co istotne, ustalony deterministycznie moment n możemy zastąpić momentem stopu, dzięki czemu otrzymujemy poniższe twierdzenie.

Twierdzenie 1.66 (Mocna własność Markowa)**.** *Niech* $(X_i)_{i=0}$ *będzie ŁM o przestrzeni stanów* $\mathcal{X}$ *i niech* τ *będzie momentem stopu względem naturalnej filtracji dla tego ŁM. Wtedy na zbiorze* $\{\tau < \infty\}$ *dla dowolnego* $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ *zachodzi*

$$\begin{aligned}\mathrm{P}\left(\{(X_{\tau+1}, X_{\tau+2}, \ldots) \in \mathcal{A}\} | X_\tau = x, \ldots, X_0\right) = \\ = \mathrm{P}\left(\{(X_1, X_2, \ldots) \in \mathcal{A}\} | X_0 = x\right)\ .\end{aligned} \tag{1.105}$$

Dowód powyższego twierdzenia znaleźć można np. w [19].

Przejdziemy teraz do prezentacji tzw. twierdzeń ergodycznych. Ich wspólnym celem jest charakteryzacja granicznego zachowania się ŁM. W literaturze (patrz np. [4, 19, 30]) twierdzenia te bywają przytaczane w różnych formach i z różnymi, choć wzajemnie sobie odpowiadającymi tezami. Z tego względu w niniejszej pracy przedstawimy je w kilku wariantach.

Twierdzenie 1.67 (Słabe twierdzenie ergodyczne dla ŁM)**.** *Niech* $(X_i)_{i=0}$ *będzie ŁM,* $\mathcal{A} \in \mathbb{B}(\mathcal{X})$ *i niech*

$$\eta_{\mathcal{A}}(n) = \frac{\mathbb{1}(X_0 \in \mathcal{A}) + \ldots + \mathbb{1}(X_n \in \mathcal{A})}{n+1} \ . \tag{1.106}$$

Wtedy dla dowolnego rozkładu początkowego lub stanu początkowego x_0

$$\mathbb{E}\left(\eta_{\mathcal{A}}(n) | X_0 = x_0\right) \xrightarrow[n\to\infty]{} \int_{x\in\mathcal{A}} d\pi_X(x) \tag{1.107}$$

i

$$\eta_{\mathcal{A}}(n) \xrightarrow[n\to\infty]{p.n.} \int_{x\in\mathcal{A}} d\pi_X(x) \ . \tag{1.108}$$

Dowód powyższego twierdzenia znaleźć można np. w [19] (dla przypadku dyskretnego) i w [30] (dla przypadku ciągłego).

Wniosek 1.68 (Ze słabego tw. ergodycznego dla ŁM)**.** *Niech* $(X_i)_{i=0}$ *będzie ŁM i niech* $f : \mathcal{X} \to \mathbb{R}$ *spełnia warunek skończoności wartości oczekiwanej*

$$\mathbb{E}_{\pi_X} |f(X)| = \int_{\mathcal{X}} |f(x)| d\pi_X(x) < \infty \ . \tag{1.109}$$

Wtedy dla dowolnego rozkładu początkowego lub stanu początkowego x_0

$$\frac{1}{n+1} \sum_{k=0}^{n} f(X_k) \xrightarrow[n\to\infty]{p.n.} \int_{\mathcal{X}} f(x)\ d\pi_X(x) \ . \tag{1.110}$$

Dowód powyższego wniosku znaleźć można np. w [4], rozdział 3.4 (dla przypadku dyskretnego) i w [30] (dla przypadku ciągłego).

Jak łatwo zauważyć, powyższe twierdzenie wraz z wnioskami są odpowiednikiem Mocnego prawa wielkich liczb (MPWL) dla zmiennych losowych *iid.* Pełnią one istotną rolę w metodzie MCMC, gdyż dają teoretyczną podstawę do estymacji wartości oczekiwanej funkcji $f(.)$ na podstawie średniej z próbki dla ŁM. Dalsze szczegóły na ten temat znaleźć można w rozdziale 6.

Twierdzenie 1.69 (Twierdzenie ergodyczne dla ŁM)**.** *Niech* $(X_i)_{i=0}$ *będzie ŁM. Wtedy dla dowolnych stanów* $x, y \in \mathcal{X}$ *mamy*

$$\lim_{n\to\infty} \mathrm{P}(X_n = y | X_0 = x) = \pi_X(y) \ . \tag{1.111}$$

Dowód powyższego twierdzenia znaleźć można np. w [19], rozdział 12.5.

Twierdzenie to podaje jeszcze inną, choć w rzeczywistości tożsamą z twierdzeniem 1.67 i wnioskiem 1.68 graniczną charakteryzację ŁM. Jak wynika z (1.111), granicznym rozkładem prawdopodobieństw przejść pomiędzy dowolnymi stanami w ŁM jest rozkład stacjonarny.

Definicja 1.70. *Zbiór $\mathcal{A} \subset \mathcal{X}$ nazwiemy* atomem*, jeśli istnieje rozkład prawdopodobieństwa $\nu(.)$, t.ż.*

$$\mathrm{P}(X_{k+1} \in \mathcal{B} | X_k = x) = \nu(\mathcal{B}) \tag{1.112}$$

dla każdego $x \in \mathcal{A}$ i każdego $\mathcal{B} \in \mathbb{B}(\mathcal{X})$.

Jak łatwo zauważyć, atom jest w pewien sposób „uogólnieniem" pojęcia „pojedynczego stanu" $x \in \mathcal{X}$ – prawdopodobieństwa przejścia do następnego stanu nie zależą od tego, w którym ze stanów atomu się znajdujemy.

Definicja 1.71. *Zbiór $\mathcal{A}$ nazwiemy* zbiorem odnowy*, jeśli istnieje stała rzeczywista $0 < \epsilon < 1$ i rozkład prawdopodobieństwa $\nu(.)$, t.ż.*

$$\mathrm{P}(X_{k+1} \in \mathcal{B} | X_k = x) \geqslant \epsilon \nu(\mathcal{B}) \tag{1.113}$$

dla każdego $x \in \mathcal{A}$ i każdego $\mathcal{B} \in \mathbb{B}(\mathcal{X})$.

Różnica pomiędzy wzorami (1.112) i (1.113) w powyższych definicjach sprowadza się do następującej własności – w przypadku atomu prawdopodobieństwo przejścia do następnego stanu jest stałe dla wszystkich stanów w atomie, natomiast w przypadku zbioru odnowy prawdopodobieństwo przejścia możemy tylko nieostro ograniczyć od dołu wykorzystując pewien stały dla tych stanów parametr ϵ.

Twierdzenie 1.72. *Niech $(X_i)_{i=0}$ będzie ŁM posiadającym atom $\mathcal{A}$ spełniający własności*

$$\mathbb{E}_{X_0 \in \mathcal{A}} \left(T_{\mathcal{A}}^2 \right) < \infty \ , \ \mathbb{E}_{X_0 \in \mathcal{A}} \left(\sum_{k=0}^{T_{\mathcal{A}}} |f(X_k)| \right)^2 < \infty \ , \tag{1.114}$$

gdzie $T_{\mathcal{A}} = \inf_{k=1,\ldots} \{X_k \in \mathcal{A}\}$ i niech

$$\sigma^2(f) = \pi_Y(\mathcal{A}) \ \mathbb{E}_{X_0 \in \mathcal{A}} \left(\sum_{k=0}^{T_{\mathcal{A}}} (f(X_k) - \mathbb{E}_{\pi_X} f(X)) \right)^2 > 0 \ , \tag{1.115}$$

wtedy

$$\frac{1}{\sqrt{n+1}} \left(\sum_{k=0}^{n} (f(X_k) - \mathbb{E}_{\pi} f(X)) \right) \xrightarrow{\mathcal{D}} N(0; \sigma^2(f)) \ . \tag{1.116}$$

Powyższe twierdzenie jest odpowiednikiem Centralnego Twierdzenia Granicznego dla przypadku zmiennych *iid*. Należy zauważyć, że zbieżność według dystrybuanty zagwarantowana jest przy stosunkowo silnym warunku na istnienie atomu $\mathcal{A}$, który dodatkowo spełnia warunki skończoności wartości oczekiwanej dla kwadratu czasu powrotu do wskazanego atomu i skończoności wartości oczekiwanej kwadratu sumy wartości zmiennych $h(X_k)$ pomiędzy wyjściem a powrotem do atomu – (1.114).

Założenie o istnieniu atomu $\mathcal{A}$ jest szczególnie istotne dla ŁM określonych na przestrzeni ciągłej. W przypadku dyskretnych ŁM dowolny stan x jest jednocześnie atomem – bowiem dla każdego $\mathcal{A} = x \in \mathcal{X}$ spełniony jest w oczywisty sposób warunek (1.112) w definicji 1.70.

Istnieją również inne wersje odpowiednika Centralnego Twierdzenia Granicznego, np. dla odwracalnych ŁM.

Definicja 1.73. *ŁM* $(X_i)_{i=0}$ *nazwiemy* odwracalnym*, jeśli zmienna* $X_{n+1}|X_n = x$ *ma ten sam rozkład co zmienna* $X_n|X_{n+1} = x$*.*

Innymi słowy, odwracalny ŁM posiada własność *odwracalności czasu* – zamiana kolejności kroków w ciągu nie wpływa na prawdopodobieństwa przejść pomiędzy krokami.

Twierdzenie 1.74. *Niech* $(X_i)_{i=0}$ *będzie odwracalnym ŁM, dla którego jest spełniony warunek*

$$0 < \sigma^2(f) := \mathrm{Var}_{\pi_X}(f(X)) + 2\sum_{k=0}^{\infty} \mathrm{Cov}_{\pi_X}\left(f(X_0), f(X_k)\right) < \infty \ . \qquad (1.117)$$

Wtedy teza (1.116) *zachodzi.*

Dowody dwóch ostatnich twierdzeń znaleźć można w [14] i [20].

Jak widzieliśmy, w powyższych twierdzeniach bardzo ważną rolę odgrywał rozkład stacjonarny $\pi_{\mathcal{X}}$, pełniąc w pewnym sensie rolę „rozkładu granicznego". Pojęciem ogólniejszym w teorii ŁM od rozkładu stacjonarnego jest *miara niezmiennicza*. Spełnia one te same warunki – (1.74) lub (1.93), co rozkład stacjonarny, ale jako *miara* nie musi być *rozkładem*, tzn. nie musi się sumować (całkować) na przestrzeni stanów $\mathcal{X}$ do jedynki. W celu uniknięcia patologii rozważać będziemy zawsze tylko σ-skończone miary niezmiennicze.

Jak się okazuje, dla danego ŁM istnieje tylko jedna miara niezmiennicza z dokładnością do stałej multiplikatywnej. Mówi o tym następujące twierdzenie.

Twierdzenie 1.75. *Jeśli* $(X_i)_{i=0}$ *jest powracającym ŁM, to istnieje dla niego* σ*-skończona miara niezmiennicza, która jest jedyna z dokładnością do stałej multiplikatywnej.*

Należy pamiętać, iż w powyższych twierdzeniach bardzo istotną rolę pełnią założenia o własnościach ŁM, co do których stosowania wprowadziliśmy umowę pod koniec rozdziałów 1.6.1 i 1.6.2. Zgodnie z tą umową, wszystkie ŁM którymi zajmujemy się w tej pracy są zawsze jednorodne, nieprzywiedlne, nieokresowe, powracalne w sensie Harrisa, posiadają rozkład stacjonarny i macierz przejścia (lub jądro przejścia).

Rozdział 2

Generatory o rozkładzie jednostajnym

2.1 Generatory fizyczne

Generatory fizyczne są to urządzenia, które generują liczby losowe wykorzystując zjawiska fizyczne. Przykładem są tu np. rzuty monetą, wyciąganie kart, itp. Generatory takie nie są jednak stosowane w praktyce. Choć bowiem ich budowa jest możliwa (np. poprzez skontruowanie urządzenia wykorzystującego zjawisko promieniotwórczości lub szumów elektronicznych), istnieją poważne problemy ze stabilnością próbek liczb losowych generowanych przez takie urządzenia. Zmiana właściwości fizycznych w otoczeniu takiego urządzenia lub samego urządzenia prowadzić może do nieprzewidywalnych zmian w generowanych ciągach. Dlatego niezbędne jest odpowiednie kalibrowanie takich urządzeń i ich częsta kontrola, co utrudnia z kolei budowę generatorów fizycznych.

Obecnie wykorzystywane są pewne namiastki generatorów fizycznych – **generatory sprzętowe** takie jak **zegar systemowy** w komputerze. Służą one jednak głównie do **inicjalizacji** (tzw. *seedowania*) najczęściej wykorzystywanych **generatorów programowych**, którymi będziemy się dalej wyłącznie zajmować.

2.2 Generatory programowe – podstawowe pojęcia

Historycznie jednym z najwcześniejszych generatorów programowych był **algorytm kwadratowy von Neumanna**. W wyniku jego działania generowane są zawsze liczby m-cyfrowe, przy czym m jest parzyste. Algorytm ten jest następujący:

1. Weź wylosowaną wcześniej liczbę losową X_{n-1}
2. Oblicz $Y_n = X_{n-1}^2$
3. Jeśli to potrzebne, dopisz odpowiednią liczbę zer na początku Y_n, tak, aby otrzymać liczbę $2m$-cyfrową

4. Za X_n przyjmij m środkowych cyfr zmodyfikowanej liczby Y_n

Przykład 2.1. *Niech $m = 2$. Dla $X_0 = 12$ znajdź ciąg wygenerowany algorytmem von Neumanna.*

Rozwiązanie: Mamy: $Y_1 = 0144, X_1 = 14, Y_2 = 0196, X_2 = 19, Y_3 = 0361, X_3 = 36, Y_4 = 1296, X_4 = 29, \ldots$ ◇

Innymi słowy, algorytm von Neumanna można zapisać w postaci

$$X_n = f(X_{n-1}) \ , \tag{2.1}$$

gdzie $f(.)$ jest pewną **ściśle określoną** funkcją matematyczną. Jak łatwo zauważyć, niezbędne jest zainicjalizowanie takiego algorytmu pewną, wybraną „ręcznie" lub wylosowaną w inny sposób, wartością X_0. Wartość ta zwana jest z j. angielskiego *seedem* i do jej wygenerowania służyć może np. wspomniany wcześniej zegar systemowy.

Należy pamiętać o pewnych ograniczeniach algorytmów określonych wzorem (2.1). Jest nim zjawisko **okresowości**. Otóż dla ciągu losowego

$$X_0, X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_{n+p}, X_{n+p+1}, \ldots \tag{2.2}$$

zawsze istnieją takie liczby n i p, że $X_i = X_{i+jp}$ dla $i \geqslant n$ i $j = 1, 2, \ldots$. Zapis ten oznacza, że co pewien czas wszystkie wylosowane liczby *powtarzają się* w dokładnie takim samym układzie złożonym z p elementów. W oczywisty sposób przy konstrukcji generatora staramy się, aby p, zwane **okresem ciągu**, było jak największe. Za w pełni losowe możemy bowiem uważać tylko liczby z ciągu $X_0, X_1, \ldots, X_{n+p}$, zwanego **okresem aperiodyczności ciągu**.

Przykład 2.2. *Wyznacz okres (poprzez obserwację!) algorytmu von Neumanna dla $X_0 = 12$.*

2.3 Generatory liniowe

Generatorem liniowym nazywamy generator opisany funkcją

$$X_{n+1} = (a_1 X_n + a_2 X_{n-1} + \ldots + a_k X_{n-k+1} + c) \mod m \tag{2.3}$$

gdzie $a_1, a_2, \ldots, a_k, c, m$ są liczbami całkowitymi, zwanymi *parametrami generatora*. Jeśli stała $c = 0$, to mówimy o **generatorze multiplikatywnym**, a dla $c \neq 0$ mówimy o **generatorze mieszanym**.

Uwaga! Dla generatora postaci (2.3) zmiennymi inicjalizującymi jego działanie są $X_0, X_1, \ldots, X_{k-1}$, czyli k zmiennych w porównaniu do jednej zmiennej wymaganej przy algorytmie von Neumanna.

Najprostszy generator liniowy ma postać

$$X_{n+1} = (aX_n + c) \mod m \tag{2.4}$$

(zaproponowany przez Lehmera). Generatory takie są często używane w standardowych bibliotekach i kompilatorach.

Przykład 2.3. *Niech $a = 2, c = 1, m = 5$. Dla $X_0 = 2$ znajdź kolejne liczby generowane przez funkcję* (2.4).

Rozwiązanie: Mamy $X_1 = (4+1) \mod 5 = 0, X_2 = (0+1) \mod 5 = 1, X_3 = 3 \mod 5 = 3, X_4 = 7 \mod 5 = 2, X_5 = 5 \mod 5 = 0, \ldots$. Zauważmy, że przy tak dobranych stałych ten generator ma okres równy zaledwie cztery! $\diamond$

Generatory liniowe można uogólnić na przypadek wielowymiarowy. Są to tak zwane **ogólne generatory liniowe**. Opisać je można przy pomocy funkcji wielu zmiennych

$$\mathbf{X}_{n+1} = \mathbf{A}\mathbf{X}_n \mod m \,, \tag{2.5}$$

gdzie $\mathbf{X}_0, \mathbf{X}_1, \ldots$ są wektorami w $\mathbb{R}^k$, zaś $\mathbf{A}$ jest macierzą $k \times k$. Powiemy o nich więcej przy zagadnieniach losowania wielowymiarowego.

2.4 Problemy z generatorami – jaki jest „odpowiedni generator"?

Wykorzystanie programowych generatorów wymaga jednak uporania się z pewnymi istotnymi problemami natury statystyczno – matematycznej:

1. Okresem generatora
2. Strukturą przestrzenną generowanych zmiennych
3. Spełnieniem warunków niezależności i pochodzenia z ustalonego rozkładu statystycznego

2.5 Okres i struktura przestrzenna

Jak zostało wspomniane już wcześniej, dla każdego generatora programowego możemy znaleźć takie wartości n i p

$$X_0, X_1, X_2, \ldots, X_n, X_{n+1}, \ldots, X_{n+p}, X_{n+p+1}, \ldots \,, \tag{2.6}$$

dla których wylosowane zmienne będą się powtarzać.

W przypadku najprostszego generatora (2.4) okres p jest równy $\min\{i > 0 : X_i = X_0\}$ i w oczywisty sposób jest mniejszy lub równy m. Co gorsze, w wielu przypadkach p jest istotnie mniejsze niż maksymalna możliwa wartość, czyli m. Istnieje szereg twierdzeń, które dla założonych parametrów generatora (2.4) określają długość osiąganego okresu. Podamy kilka przykładów takich lematów.

Lemat 2.4. *Generator multiplikatywny*

$$X_{n+1} = (aX_n) \mod m \tag{2.7}$$

ze stałą $m = 2^L$ dla $L \geqslant 4$ osiąga maksymalny okres równy 2^{L-2} wtedy i tylko wtedy, gdy X_0 jest liczbą nieparzystą oraz $a = 3 \mod 8$ (czyli a jest podzielne przez 8 z resztą 3) lub $a = 5 \mod 8$.

Przykład 2.5. *Generator RANDU: $a = 2^{16} + 3, m = 2^{31}$.*

Lemat 2.6. *Generator mieszany* (2.4) *osiąga pełny okres m wtedy i tylko wtedy, gdy są spełnione wszystkie trzy warunki:*

1. *liczby c i m nie mają wspólnych dzielników,*

2. $a = 1 \mod p$ *dla każdego czynnika pierwszego liczby* m *(czyli każdej liczby pierwszej* p *dzielącej* m*),*

3. $a = 1 \mod 4$*, jeżeli* 4 *jest dzielnikiem liczby* m*.*

Przykład 2.7. *Niech* $a = 69069, c = 1, m = 2^{32}$*.*

Przykład 2.8. *Na podstawie powyższych twierdzeń stwórz własne przykłady odpowiednich generatorów.*

Należy podkreślić, że wybór odpowiednio dużej wartości m jest bardzo istotny – generator liniowy (2.4) nie może nigdy „wyprodukować" więcej niż m liczb losowych, bo istnieje tylko m różnych reszt z operacji modulo m. Ponieważ zazwyczaj interesują nas liczby rozłożone równomiernie na przedziale $[0; 1]$, wygenerowane wartości $X_1, X_2, \ldots$ modyfikujemy, obliczając

$$U_i = \frac{X_i}{m} \ . \tag{2.8}$$

Również to przekształcenie wskazuje na wybór odpowiednio dużej wartości m – liczby (punkty) $U_1, U_2, \ldots$ powinny dostatecznie „gęsto" pokrywać przedział $[0; 1]$. Przedział $[0; 1]$ będziemy dalej w skrócie określać *przedziałem jednostkowym*. Z kolei o $[0; 1]^l$ dla $l > 1$ będziemy mówić jako o *l-wymiarowej kostce jednostkowej*.

Drugim problemem jest struktura przestrzenna wylosowanych zmiennych. Dla ustalonego $d \geqslant 1$ rozważmy punkty w d-wymiarowej przestrzeni, opisane współrzędnymi

$$(U_1, U_2, \ldots, U_d), (U_2, U_3, \ldots, U_{d+1}), \ldots \tag{2.9}$$

Jeśli ciąg $U_1, U_2, \ldots$ jest „dostatecznie losowy", to wspomniane punkty (2.9) powinny być „równomiernie" i „przypadkowo" rozłożone w d-wymiarowej kostce jednostkowej. Tymczasem w przypadku generatorów liniowych punkty te często układają się w regularne, geometryczne (czyli zupełnie „nielosowe") struktury, np. pionowe lub poziome „pasy", „pasy na krzyż", itd.

Oprócz współrzędnych (2.9) można analizować także inne zestawy, np.

$$(U_1, U_2, \ldots, U_d), (U_{d+1}, U_3, \ldots, U_{2d}), \ldots \tag{2.10}$$

i badać zachowanie się takich punktów przy wypełnianiu kostki jednostkowej $\mathbb{I}^d$.

W celu pomiaru zagęszczenia punktów w $\mathbb{I}^d$ stosuje się specjalne miary, np. *test spektralny*.

2.6 Testy statystyczne dla generatorów

Należy pamiętać, że wszelkie zmienne wygenerowane za pomocą generatorów programowych tak naprawdę są tylko liczbami *pseudolosowymi*. Z samej swej konstrukcji (zastosowanie deterministycznych funkcji w celu określenia kolejnej wartości X_{n+1}) są ciągami *deterministycznymi*, a nie *losowymi*. Jeżeli chcemy ciąg $(X_i)_{i=1}$ wykorzystać jako „losowy" musi on spełniać prosty warunek – dostatecznie dobrze naśladować „losowość".

Co to znaczy? Przypuśćmy, że chcemy, aby nasz generator tworzył ciąg zmiennych niezależnych z rozkładu równomiernego na przedziale jednostkowym. Z samego tego opisu możemy wyciągnąć wnioski dotyczące niezbędnych do spełnienia warunków:

- Ciąg $(X_i)_{i=1}$ musi być ciągiem niezależnych (w sensie probabilistycznym) zmiennych
- Zmienne $(X_i)_{i=1}$ muszą pochodzić z rozkładu jednostajnego na przedziale jednostkowym.

W celu sprawdzenia obu (albo i więcej) warunków należy wykonać odpowiednie testy statystyczne. Np. drugi warunek implikuje, że jeśli zmienne mają pochodzić z rozkładu jednostajnego na przedziale $[0;1]$, to wartość oczekiwana utworzonego ciągu musi być równa 0,5, a wariancja musi wynosić $\frac{1}{12}$. Sprawdzić też możemy np. wygląd histogramu (powinien być dostatecznie „równomierny"), przeprowadzić test zgodności z rozkładem jednostajnym, itp. Przyjrzyjmy się przykładom różnych podejść.

Niezależność poszczególnych zmiennych w ciągu

$$X_1, X_2, X_3, \ldots \tag{2.11}$$

może zostać sprawdzona przy pomocy najprostszego z narzędzi, czyli współczynnika korelacji liniowej

$$\mathrm{Corr}(Y,Z) = \frac{\mathbb{E}(Y-\mathbb{E}\,Y)(Z-\mathbb{E}\,Z)}{DY\ DZ}\ . \tag{2.12}$$

Jak wiadomo, jeśli wartość tego współczynnika jest bliska zeru, to zmienne Y i Z są niezależne liniowo (ale niekoniecznie w ogóle niezależne). Aby móc zastosować obliczanie współczynnika korelacji i testy oparte na tym współczynniku, niezbędne jest podzielenie badanego ciągu $(X_i)_{i=1}$ na dwie „podgrupy" (grające role zmiennych Z i Y odpowiednio). Można np. podzielić ciąg $(X_i)_{i=1}$ na dwie połowy i zastosować wtedy wzór

$$\rho = \frac{\frac{1}{n/2}\sum_{i=1}^{n/2}\left(X_i-\bar{X}_1\right)\left(X_{n/2+i}-\bar{X}_2\right)}{\sqrt{\frac{1}{n/2}\sum_{i=1}^{n/2}\left(X_i-\bar{X}_1\right)^2\frac{1}{n/2}\sum_{i=n/2+1}^{n}\left(X_i-\bar{X}_2\right)^2}}\ , \tag{2.13}$$

gdzie $\bar{X}_1$ jest średnią dla pierwszej połowy ciągu $(X_i)_{i=1}$, a $\bar{X}_2$ – dla drugiej. Założyliśmy przy tym dla ułatwienia zapisu, że n jest liczbą parzystą.

Na ciąg $(X_i)_{i=1}$ możemy również popatrzeć jak na szereg czasowy i wykorzystać narzędzia analizy szeregów czasowych. W szczególności możemy badać model postaci

$$X_n = aX_{n-1} + b \tag{2.14}$$

i sprawdzić wartość współczynników autokorelacji poszczególnych rzędów (tzn. współczynniki korelacji dla par (X_n, X_{n-1}), (X_n, X_{n-2}), itd.). Jak się łatwo domyślić, wartości tych współczynników powinny być bliskie zeru, aby móc w ogóle mówić o niezależności.

W celu sprawdzenia, czy ciąg $(X_i)_{i=1}$ pochodzi z ustalonego przez nas rozkładu należy również zastosować odpowiednie testy statystyczne. Najprostszym przykładem może być tutaj test χ^2 zgodności. W teście tym dzielimy całą przestrzeń wartości zmiennej losowej na pewną liczbę przedziałów, np. przedział $[0;1]$ na m podprzedziałów. Jeśli nasze obserwacje pochodzą z rozkładu jednostajnego na $[0;1]$, to w każdym podprzedziale (długości $\frac{1}{m}$) powinno ich być tyle samo (dokładnie w liczbie $\frac{n}{m}$ w podprzedziale). Korzystając ze statystyki

$$\chi^2 = \sum_{j=1}^{m}\frac{\left(n_j-\frac{n}{m}\right)^2}{\frac{n}{m}}\ , \tag{2.15}$$

gdzie n_j jest liczbą punktów ciągu $(X_i)_{i=1}$ zawartych w podprzedziale j-tym, możemy sprawdzić słuszność naszego przypuszczenia. Jeśli $\chi^2 > \chi^2_{1-\alpha,m-1}$, to odrzucamy hipotezę, o tym, że punkty w badanym ciągu pochodzą z rozkładu jednostajnego, przy czym $\chi^2_{1-\alpha,m-1}$ jest kwantylem rozkładu χ^2 o rzędzie $1-\alpha$ i liczbie stopni swobody $m-1$.

Innym testem, który możemy zastosować jest np. test Kołmogorowa–Smirnowa.

Ze względu na istotne powiązanie pomiędzy badanymi przez nas algorytmami generującymi liczby losowe a testami statystycznymi, wprowadzimy poniższą „roboczą" definicję.

Definicja 2.9. *Powiemy, że dany algorytm generuje liczby (pseudo)losowe, jeśli żaden z wybranych przez nas testów statystycznych nie odrzuca hipotezy o tym, że ciąg $(X_i)_{i=1}$ wygenerowany tym algorytmem jest ciągiem zmiennych* iid *z ustalonego rozkładu statystycznego.*

Choć definicja ta jest bardzo użyteczna z praktycznego i punktu widzenia, należy podkreślić jej wady:

- uznaniowość wyboru testów statystycznych – może się zdarzyć, że wszystkie wcześniej wykorzystywane testy nie odrzucały hipotezy, iż dany algorytm generuje liczby (pseudo)losowe, ale nowy, dodatkowy test może taką hipotezę odrzucić,
- problem wyboru ciągu – może się zdarzyć, że algorytm generuje liczby (pseudo)losowe dla wszystkich przebadanych przez nas ciągów, ale pewien nowy ciąg lub nowe parametry inicjalizujące algorytm doprowadzą do odrzucenia hipotezy o tym, że dany algorytm generuje liczby (pseudo)losowe.

W związku z tym niezbędny jest wybór odpowiednio szerokiego zbioru testów statystycznych (jak np. testy w kolekcji „Die Hard") oraz sprawdzenie szerokiego spektrum ciągów i parametrów seedowania. Inne podejście wymaga *ścisłego matematycznego dowodu*, że algorytm generuje odpowiednie liczby losowe. Choć jest ono ściślejsze i „pewniejsze" w wielu przypadkach odpowiednie dowody nie są znane.

2.7 Generatory Fibonacciego

Nazwisko Fibonacciego związane jest ze słynnym rekurencyjnym ciągiem

$$X_n = X_{n-1} + X_{n-2} \text{ dla } n \geqslant 2 \ , \tag{2.16}$$

gdzie $X_0 = X_1 = 1$. Podobny do niego ciąg

$$X_n = (X_{n-1} + X_{n-2}) \mod m \text{ dla } n \geqslant 2 \tag{2.17}$$

generuje liczby, które można w pewnym przybliżeniu uznać za losowe. Testy statystyczne nie odrzucają hipotezy, iż ciąg $(X_i)_{i=1}$ wygenerowany (2.17) pochodzi z rozkładu jednostajnego, ale odrzucają hipotezę, iż ciąg ten jest ciągiem zmiennych niezależnych. W związku z tym uogólniono wzór (2.17) do postaci

$$X_n = (X_{n-s} + X_{n-r}) \mod m \text{ dla } n \geqslant r, r > s \geqslant 1 \ . \tag{2.18}$$

Jest to tzw. ALFG (Additive Lagged Fibonacci Generator). W przeszłości takie generatory były jednak rzadko stosowane, ze względu na swoją mniejszą szybkość od generatorów multiplikatywnych i konieczność przechowywania większej liczby wcześniej wygenerowanych wartości. Obecnie jednak, dzięki szybkim komputerom i tańszej pamięci, są znacznie częściej spotykane. Co więcej, mogą one mieć większy okres niż generatory multiplikatywne. Dla $m = 2^L$ maksymalny okres ALFG wynosi $(2^r - 1)2^{L-1}$.

Generatory postaci (2.18) można uogólnić do wzoru

$$X_n = (X_{n-s} \diamond X_{n-r}) \mod m \ , \tag{2.19}$$

gdzie $\diamond$ jest pewną operacją (np. dodawaniem, odejmowaniem, mnożeniem). W przypadku mnożenia mówimy o MLFG (Multiplicative Lagged Fibonacci Generator), a jego maksymalny okres wynosi $(2^r - 1)2^{L-3}$.

Przykładem uogólnienia może też być zastosowanie operacji bitowej *xor* (Two-tap Generalised Feedback Shift Register), co daje algorytm związany z generatorami opartymi na rejestrach przesuwnych.

2.8 Łączenie generatorów

Bardzo dobre wyniki daje *kombinowanie* generatorów, czyli odpowiednie łączenie dwóch lub większej liczby prostszych generatorów.

Załóżmy mianowicie, że dysponujemy dwoma generatorami, które wygenerowały odpowiednio ciągi $(X_i)_{i=1}$ i $(Y_i)_{i=1}$. Wtedy możemy stworzyć nowy ciąg według wzoru

$$Z_i = X_i \diamond Y_i \ , \tag{2.20}$$

gdzie $\diamond$ jest znowu pewnym działaniem (np. dodawaniem, odejmowaniem, operacją xor, działaniem modulo). Dzięki takiej operacji zazwyczaj ciąg $(Z_i)_{i=1}$ ma lepsze własności niż ciągi go tworzące – jest „bardziej równomierny", „bardziej niezależny", często ma też większy okres. Jeśli np. ciąg $(X_i)_{i=1}$ ma okres p_1, a $(Y_i)_{i=1}$ ma okres p_2, przy czym p_1 i p_2 są względnie pierwsze, to $(Z_i)_{i=1}$ ma okres $p_1 p_2$ (z tzw. chińskiego twierdzenia o resztach).

2.9 Generatory nieliniowe

Rozważane wcześniej generatory były oparte na liniowych wzorach rekurencyjnych. Ułatwiało to ich implementację, powodowało jednak omówione już problemy ze strukturą przestrzenną generowanych punktów rozważanych dla wielowymiarowych kostek. Dlatego można wykorzystywać generatory nieliniowe, oparte np. na odwracaniu lub obliczaniu kwadratów. Przykładem może być tutaj generator postaci

$$X_n = \left(a\check{X}_{n-1}^{-1} + b\right) \mod m \ , \tag{2.21}$$

gdzie m jest liczbą pierwszą, a operacja $\check{X}^{-1}$ jest *odwracaniem modulo*. Operacja ta jest zdefiniowana następująco: dla $x = 0$ zachodzi $\check{x}^{-1} \mod m = 0$, a dla $x \neq 0$ liczba $\check{x}^{-1}$ musi spełniać warunek $x \cdot \check{x}^{-1} \mod m = 1$.

Przykład 2.10. *Jakie wartości otrzymujemy dla generatora odwracania modulo o postaci*

$$X_n = \left(2\check{X}_{n-1}^{-1} + 3\right) \mod 7 \tag{2.22}$$

zainicjowanego wartością $X_0 = 2$*?*

Rozwiązanie: Z definicji odwracania modulo mamy

X	0	1	2	3	4	5	6
$\check{X}^{-1}$	0	1	4	5	2	3	6

stąd

$$X_1 = (2 \cdot 4 + 3) \mod 7 = 4\ ,\ X_2 = (2 \cdot 2 + 3) \mod 7 = 0\ ,$$
$$X_3 = (2 \cdot 0 + 3) \mod 7 = 3, \ldots \tag{2.23}$$

◇

Przykładem może też być generator oparty na obliczaniu kwadratów, opisany wzorem

$$X_n = X_{n-1}^2 \mod m\ . \tag{2.24}$$

Rozdział 3

Generatory o różnych rozkładach prawdopodobieństwa

Należy podkreślić, że generowanie zmiennych losowych o dowolnym rozkładzie, innym niż równomierny, przebiega zazwyczaj w dwóch etapach. Pierwszym z nich jest wygenerowanie zmiennej U o rozkładzie równomiernym, drugim – zastosowanie odpowiedniego algorytmu w celu uzyskania zmiennej X o innym rozkładzie (np. normalnym standardowym) przy wykorzystaniu wartości U.

Obecnie skupimy się na opisaniu algorytmów i metod wykorzystywanych w tym drugim etapie. Zakładać więc będziemy, że dysponujemy już odpowiednim generatorem zmiennych losowych o rozkładzie jednostajnym na przedziale jednostkowym $[0; 1]$. Dla ułatwienia notacji i przejrzystości algorytmów, przypuśćmy, że mamy funkcję `GenerujU`, której kolejne wywołania produkują ciąg zmiennych *iid* z rozkładu $U[0; 1]$.

Metody generowania zmiennej losowej o rozkładzie innym niż rozkład równomierny można podzielić na dwie grupy: metody ogólne (do zastosowania dla wielu różnych rozkładów) oraz metody „specyficzne" (generujące tylko zmienne ze ściśle określonych rozkładów, najczęściej na podstawie bardzo wysublimowanych i „sztuczkowych" algorytmów). Rozpoczniemy od przeglądu metod należących do pierwszej grupy.

3.1 Metoda odwracania dystrybuanty

W metodzie tej wykorzystujemy następujące twierdzenie.

Twierdzenie 3.1. *Niech U będzie zmienną losową z rozkładu jednostajnego na przedziale jednostkowym, a $F_X(.)$ – ciągłą i ściśle rosnącą dystrybuantą pewnego rozkładu prawdopodobieństwa. Wtedy zmienna losowa X określona warunkiem*

$$X = F_X^{-1}(U) \tag{3.1}$$

ma rozkład dany dystrybuantą $F_X(.)$.

Dowód. Ponieważ $F_X(.)$ jest funkcją ciągłą i ściśle rosnącą, zatem istnieje funkcja odwrotna $F_X^{-1}(.)$. Dla (3.1) mamy

$$\mathrm{P}(X \leqslant x) = \mathrm{P}\left(F_X^{-1}(U) \leqslant x\right) = \mathrm{P}\left(U \leqslant F_X(x)\right) = F_X(x) \ , \tag{3.2}$$

zatem zmienna X ma rozkład zadany dystrybuantą $F_X(.)$. □

Powyższe twierdzenie prowadzi bezpośrednio do następującego algorytmu

Algorytm 3.2.

```
U = GenerujU
X = F^(-1) (U)
return X
```

gdzie `F^(-1) (.)` jest wartością funkcji $F_X^{-1}(.)$. Jak widzimy, jeśli dla określonego rozkładu prawdopodobieństwa możemy znaleźć funkcję $F_X^{-1}(.)$ odwrotną do dystrybuanty tego rozkładu, to algorytm ten umożliwia generowanie zmiennej o tym rozkładzie.

Przykład 3.3. *Przypuśćmy, że zmienna X ma pochodzić z rozkładu jednostajnego na przedziale $[a;b]$. Ponieważ gęstość takiego rozkładu wynosi*

$$f(t) = \frac{1}{b-a} \ , \tag{3.3}$$

zatem dystrybuanta rozkładu $U([a;b])$ wyraża się wzorem

$$F(x) = \frac{x-a}{b-a} \ . \tag{3.4}$$

Ponieważ $F(.)$ jest ciągła i ściśle rosnąca na przedziale $[a;b]$, zatem funkcja odwrotna do dystrybunaty istnieje i jest postaci

$$F^{-1}(y) = y(b-a) + a \ . \tag{3.5}$$

Zatem z tw. 3.1 zmienna losowa

$$X = U(b-a) + a \tag{3.6}$$

posiada zgodnie z naszym wymaganiem rozkład jednostajny na przedziale $[a;b]$. Odpowiedni algorytm ma zatem postać

Algorytm 3.4.

```
U = GenerujU
X = U * (b-a) + a
return X
```

Przykład 3.5. *Jeśli zmienna X ma pochodzić z rozkładu wykładniczego, to funkcja odwrotna do dystrybuanty ma postać*

$$F^{-1}(y) = -\frac{1}{\lambda}\ln(1-y) \ , \tag{3.7}$$

stąd zmienna

$$X = -\frac{1}{\lambda}\ln U \tag{3.8}$$

posiada oczekiwany rozkład.

Twierdzenie 3.1 można rozszerzyć na przypadek dowolnych, niekoniecznie ciągłych i ściśle rosnących dystrybuant $F_X(.)$. Jak wiadomo, w takich przypadkach funkcja odwrotna nie musi istnieć. Wystarczy jednak posłużyć się *uogólnioną definicją funkcji odwrotnej do dystrybuanty*. Niech

$$F_X^-(t) = \inf\{x : t \leqslant F(x)\} \ . \tag{3.9}$$

Zgodnie z tą definicją, $F_X^{-1}(t)$ jest to najmniejsza wartość funkcji dystrybuanty $F_X(x)$, która przekracza lub jest równa t.

Twierdzenie 3.6. *Niech U będzie zmienną losową z rozkładu jednostajnego na przedziale jednostkowym, a $F_X(.)$ dystrybuantą pewnego rozkładu prawdopodobieństwa. Wtedy zmienna losowa X zdefiniowana warunkiem*

$$X = F_X^-(U) \tag{3.10}$$

ma rozkład określony dystrybuantą $F_X(.)$.

Przykład 3.7. *Jeśli zmienna X ma pochodzić z rozkładu dyskretnego określonego ciągiem prawdopodobieństw p_i, tzn. $p_i = \mathrm{P}(X = x_i)$, to powinniśmy zastosować przekształcenie*

$$X = \min\left\{k : U \leqslant \sum_{i=1}^{k} p_i\right\} \ . \tag{3.11}$$

Należy podkreślić, że pomimo ogólności wyżej przytoczonych twierdzeń, bazujące na nich algorytmy, poza np. generatorem dla rozkładu wykładniczego, raczej rzadko stosuje się w praktyce. Trudność wynika głównie z faktu stosowania funkcji odwrotnej do dystrybuanty, co w wielu przypadkach wymagałoby skomplikowanego numerycznie podejścia. W związku z tym stosuje się często szybsze i prostsze metody, które omówimy w kolejnych rozdziałach.

Przykład 3.8. *Dla rozkładu normalnego standardowego $N(0,1)$ zastosować można przybliżenie następującą funkcją odwrotną*

$$F_X^{-1}(t) = \begin{cases} g(t) & \text{dla } 10^{-20} < t < 0,5 \\ -g(1-t) & \text{dla } 0,5 \leqslant t < 1 - 10^{20} \end{cases} \ , \tag{3.12}$$

gdzie

$$g(t) = \sqrt{-2\ln t} - \frac{L(\sqrt{-2\ln t})}{M(\sqrt{-2\ln t})} \tag{3.13}$$

i $L(.)$ oraz $M(.)$ są pewnymi określonymi wielomianami stopnia czwartego.

Należy podkreślić, że jeśli ciąg $U_1, U_2, \ldots$ jest ciągiem zmiennych *iid* z rozkładu równomiernego na przedziale jednostkowym, to uzyskane metodą odwracania dystrybuanty zmienne $X_1, X_2, \ldots$ też będą zmiennymi *iid*. Wynika to z odpowiedniego twierdzenia z teorii rachunku prawdopodobieństwa. Zauważmy ponadto, że w opisywanej metodzie do wygenerowania jednej zmiennej X_i wystarczy tylko jedna zmienna U_i. Jest to jej ogromną zaletą.

3.2 Metoda eliminacji

Rozpatrzmy najpierw uproszczoną wersję tej metody, aby potem przejść do jej ogólnej wersji.

Załóżmy, że interesuje nas wygenerowanie zmiennej losowej X o gęstości zadanej funkcją $f_X(t)$, która to funkcja jest większa od zera tylko na przedziale $[0;1]$ i równa zeru poza tym przedziałem. Przypuśćmy ponadto, że funkcja $f_X(t)$ przyjmuje na przedziale jednostkowym wartości ograniczone przez pewną stałą M. W takim przypadku następujący algorytm

Algorytm 3.9.

```
repeat
 {U1 = GenerujU;
 U2 = GenerujU }
until M * U2 <= f (U1)
X = U1
return X
```

generuje zmienną o rozkładzie zadanym gęstością $f_X(.)$. Algorytm ten tworzy punkt `(U1,MU2)` o niezależnych współrzędnych, wylosowany „równomiernie" z prostokąta $[0;1] \times [0;M]$. Jeśli punkt ten znajduje się pod wykresem funkcji gęstości $f_X(.)$, to zostaje zaakceptowany i jego współrzędna `U1` staje się zwracaną zmienną X. W przeciwnym przypadku losowanie obu współrzędnych punktu zostaje powtórzone.

Ogólny wariant tej metody zakłada, że umiemy generować zmienną Y o rozkładzie określonym funkcją gęstości $g_Y(t)$, zwaną gęstością dominującą. Załóżmy ponadto, że interesuje nas zmienna X opisana gęstością $f_X(t)$, przy czym na całym przedziale określoności tej gęstości zachodzi $f_X(t) \leqslant Mg_Y(t)$ dla pewnej stałej M. Wtedy następujący algorytm

Algorytm 3.10.

```
repeat
 {U = GenerujU
 Y = GenerujG }
until M * U * g (Y) <= f (Y)
X = Y
return X
```

generuje zmienną X z rozkładu o gęstości $f_X(.)$. Zauważmy, że w algorytmie tym wykorzystujemy funkcję `GenerujG`, która służy do generowania zmiennej z rozkładu o gęstości $g_Y(.)$. Algorytm ten ma podobną interpretację co poprzedni – generujemy punkt o współrzędnych `(Y,M*U*g (Y))`, który jeśli znajduje się pod wykresem funkcji $f_X(Y)$, to zostaje zaakceptowany i jego współrzędna `Y` staje się zwracaną zmienną X. W przeciwnym przypadku losowanie obu współrzędnych punktu zostaje powtórzone.

W obu tych algorytmach istotną rolę odgrywała stała M. Łatwo zauważyć, że jeśli zachodzi $f_X(t) \leqslant Mg_Y(t)$ dla pewnego M, to dla dowolnego $M_1 \geqslant M$ warunek ograniczający $f_X(t) \leqslant M_1 g_Y(t)$ też będzie zachodzić. Dlatego interesującym jest wyznaczenie odpowiedniej wartości stałej M. Wynika to z prostej

obserwacji, iż w algorytmie wielokrotnie jest powtarzane losowanie punktu, dopóki nie zostanie spełniony warunek akceptacji. Warunek ten zależy od stałej M. Stąd *prawdopodobieństwo* akceptacji punktu również zależy od M, przy czym wraz ze wzrostem M prawdopodobieństwo to maleje! Dlatego też powinniśmy postarać się, aby akceptacja następowała jak najszybciej, czyli by prawdopodobieńswo zajścia warunku (tzw. warunku akceptacji)

$$MUg_Y(Y) \leqslant f_X(Y) \tag{3.14}$$

było jak największe. Zauważmy, że

$$\mathrm{P}(MUg(Y) \leqslant f(Y)) = \int_{\mathcal{Y}} \int_0^{f(y)/Mg(y)} g(y)\, du\, dy = \frac{1}{M}\ , \tag{3.15}$$

zatem optymalną wartością M jest

$$M^* = \min\{M : f_X(.) \leqslant Mg_Y(.)\}\ . \tag{3.16}$$

Przykład 3.11. *Wygenerujemy zmienną X o rozkładzie normalnym standardowym. W tym celu najpierw wygenerujemy zmienną X_1 o gęstości $f_{X_1}(t) = \sqrt{2/\pi}\exp(-x^2/2)$ (czyli jest to prawa połówka rozkładu $N(0,1)$), a później przekształcimy ją do zmiennej X_2 dodając znak „+" lub „-" z równymi prawdopodobieństwami 0,5. Gęstością dominującą dla $f_{X_1}(.)$ będzie gęstość rozkładu wykładniczego* $\mathrm{Exp}(\lambda)$ *równa $g_Y(t) = e^{-t}$, z której umiemy już generować (patrz Przykład 3.5). Jak się okazuje, minimalna stała M ma wartość $\sqrt{2e/\pi}$, stąd algorytm generujący X_1 ma postać*

Algorytm 3.12.

```
repeat
 {U = GenerujU
 Y = GenerujWykladniczy }
until Sqrt(2e /pi) * U * exp(-Y) <= Sqrt(2/pi) * exp(-Y^2 / 2)
X = Y
return X
```

W drugim kroku generujemy zmienną X_2, korzystając z dodatkowej, niezależnej zmiennej losowej o rozkładzie dwupunktowym (patrz Przykład 3.7).

Algorytm 3.13.

```
U = GenerujU
if U <= 0,5 then X = - X
return X
```

Zauważmy, że w powyższym przykładzie warunek akceptacji w pierwszym kroku ma postać

$$\sqrt{2e/\pi}U\exp(-Y) \leqslant \sqrt{2/\pi}\exp(-Y^2/2)\ , \tag{3.17}$$

co można łatwo uprościć do postaci

$$\sqrt{e}U\exp(-Y) \leqslant \exp(-Y^2/2)\ . \tag{3.18}$$

Prowadzi to do istotnego wniosku: nie musimy znać pełnej postaci funkcji gęstości $f_X(.)$, a jedynie tą postać z dokładnością do *stałej normującej*. Parametr M wykorzystywany w tej metodzie „bierze pod uwagę" wszystkie stałe (o ile tylko jest dobrze wyznaczony). Fakt ten jest dodatkowo potwierdzany przez poniższy lemat, który stanowi jednocześnie dowód słuszności metody eliminacji.

Lemat 3.14. *Dla metody eliminacji w postaci ogólnej zachodzi*

$$\mathrm{P}(Y \leqslant x | Y \text{ zaakceptowany}) = \frac{\int_{-\infty}^{x} f_X(t)dt}{\int_{-\infty}^{\infty} f_X(t)dt} \ . \tag{3.19}$$

Dowód. Mamy

$$\mathrm{P}(Y \leqslant x \wedge Y \text{ zaakceptowany}) = \int_{-\infty}^{x} \frac{f_X(t)}{M g_Y(t)} g_Y(t)\ dt = \int_{-\infty}^{x} \frac{f_X(t)}{M}\ dt \ , \tag{3.20}$$

gdzie w równości wykorzystujemy gęstość warunkową akceptacji. Stąd bezpośrednio

$$\mathrm{P}(Y \text{ zaakceptowany}) = \int_{-\infty}^{\infty} \frac{f_X(t)}{M}\ dt \tag{3.21}$$

i korzystając z twierdzenia o prawdopodobieństwie warunkowym otrzymujemy

$$\mathrm{P}(Y \leqslant x | Y \text{ zaakceptowany}) = \frac{\mathrm{P}(Y \leqslant x \wedge Y \text{ zaakceptowany})}{\mathrm{P}(Y \text{ zaakceptowany})} \tag{3.22}$$

co prowadzi do tezy lematu. □

Z (3.19) widzimy jednoznacznie, że niezależnie od postaci stałej normującej, dzięki czynnikowi w mianowniku $\int_{-\infty}^{\infty} f_X(t)dt$ metoda eliminacji produkuje zmienne o odpowiednio unormowanej gęstości $f_X(.)$.

3.3 Metoda szybkiej eliminacji i szeregów

Warunek akceptacji

$$MU g_Y(Y) \leqslant f_X(Y) \tag{3.23}$$

wymaga za każdym razem obliczenia wartości funkcji gęstości $f_X(.)$ i $g_Y(.)$ w pewnym ustalonym punkcie. Może to być czasami kłopotliwe ze względów numerycznych. Dlatego też skuteczniejsze okazać się może znalezienie prostszych funkcji ograniczających, postaci

$$\alpha_1(x) \leqslant \frac{f_X(x)}{M g_Y(x)} \leqslant \beta_1(x) \tag{3.24}$$

dla dowolnego x. Łatwo zauważyć, że jeśli wylosowane w algorytmie zmienne – U z rozkładu jednostajnego i Y z rozkładu $g_Y(.)$ spełniają warunek

$$U \leqslant \alpha_1(Y) \tag{3.25}$$

(tzw. warunek szybkiej akceptacji), to warunek (3.23) też będzie spełniony, zatem zmienna Y zostanie zaakceptowana.

Jeśli z kolei zmienne U i Y będą spełniać warunek

$$U \geqslant \beta_1(Y) \tag{3.26}$$

(tzw. warunek szybkiej eliminacji), to warunek (3.23) też nie będzie spełniony, zatem zmienna Y zostanie odrzucona.

Innymi słowy, jeśli znajdziemy dostatecznie proste funkcje $\alpha_1(.)$ i $\beta_1(.)$, które „obustronnie" będą przybliżać warunek akceptacji, może to przyśpieszyć algorytm generowania zmiennej losowej. Algorytm ten ma wtedy postać

Algorytm 3.15.

```
flaga = 0
repeat
 {U = GenerujU
 Y = GenerujG
 if U <= alpha (Y) then
  flaga = 1
 else
  if U <= beta (Y) then
   if M * U * g (Y) <= f (Y) then
    flaga = 1}
until flaga = 1
X = Y
return X
```

Zagnieżdzenie warunków „if... then..." w powyższym algorytmie ma na celu jak najrzadsze wykonywanie sprawdzenia warunku (3.23).

Metodę tą można uogólnić na ciąg funkcji przybliżających warunek akceptacji. Jest to tzw. metoda szeregów. Przypuśćmy, że dla dowolnych x i $n \in \mathbb{N}$ zachodzi

$$\underline{f}_n(x) \leqslant f_X(x) \leqslant \overline{f}_n(x) \ , \tag{3.27}$$

zatem gęstość docelowa $f_X(.)$ jest przybliżana przez odpowiedni ciąg funkcji. Przykładowy algorytm ma wtedy postać

Algorytm 3.16.

```
repeat
 {U = GenerujU
 Y = GenerujG
 n = 0
 repeat
  {n = n + 1
  if M * U * g (Y) <= f_n (Y) then return Y}
 until M * U * g (Y) > f^n (Y)
until false
```

Bazując na (3.24) metodę tą możemy zapisać również jako ciąg warunków szybkiej akceptacji i szybkiego odrzucenia

$$\alpha_k(x) \leqslant \ldots \leqslant \alpha_1(x) \leqslant \frac{f_X(x)}{M g_Y(x)} \leqslant \beta_1(x) \leqslant \ldots \beta_l(x) \tag{3.28}$$

dla dowolnego x. W takim przypadku warunki sprawdza się w kolejności

$$U \leqslant \alpha_k(Y), U \geqslant \beta_l(Y), U \leqslant \alpha_{k-1}(Y), \ldots, U \leqslant \frac{f_X(Y)}{Mg_Y(Y)} \tag{3.29}$$

aż do momentu, w którym jeden z nich nie zostanie spełniony i nastąpi przyjęcie lub odrzucenie zmiennej Y.

Metoda szeregów w postaci (3.27) może zostać uogólniona na przykład szeregów zbieżnych. Przypuśćmy, że funkcja gęstości $f_X(.)$ wyraża się jako granica zbieżnego szeregu nieskończonego

$$f_X(x) = \sum_{i=1}^{\infty} S_i(x) \ , \tag{3.30}$$

przy czym potrafimy zawsze oszacować bezwzględną wartość reszty szeregu

$$\left| \sum_{i=n}^{\infty} S_i(x) \right| \leqslant R_n(x) \tag{3.31}$$

dla dowolnego x. Odpowiedni algorytm ma wtedy postać

Algorytm 3.17.

```
repeat
 {U = GenerujU
 Y = GenerujG
 S = 0
 n = 0
 repeat
  {n = n + 1
  S = S + S_n (Y)}
 until | S - M * U * g (Y) | > R_{n+1} (Y)}
until M * U * g (Y) <= S (Y)
X = Y
return X
```

Zauważmy, że wewnętrzna pętla „repeat... until..." oblicza kolejne przybliżenia funkcji $f_X(.)$ poprzez sumowanie funkcji $S_i(.)$. To sumowanie kończy się w momencie, gdy różnica pomiędzy przybliżeniem a wielkością z warunku akceptacji $MUg_Y(.)$ jest mniejsza niż pozostała reszta szeregu $R_{n+1}(.)$ (czyli $S_n(.)$ wraz z błędem $R_{n+1}(.)$ jest na pewno mniejsze lub większe od $MUg_Y(.)$). Wtedy, dysponując już odpowiednim przybliżeniem, następuje sprawdzenie warunku akceptacji.

3.4 Metoda ilorazu równomiernego

Metoda *ilorazu równomiernego* (*RI*, *ROU*, czyli *ratio-of-uniforms*) bazuje na następującym twierdzeniu.

Twierdzenie 3.18. *Niech $f_X(.)$ będzie nieujemną i skończenie całkowalną funkcją i niech*

$$\mathcal{C}_f = \left\{ (u,v) : 0 \leqslant u \leqslant \sqrt{f_X\left(\frac{v}{u}\right)} \right\} \ . \tag{3.32}$$

Jeśli punkt (U, V) ma rozkład równomierny na zbiorze $\mathcal{C}_f$, to zmienna losowa $X = \frac{V}{U}$ ma rozkład o gęstości $\frac{f_X(.)}{\int f_X(t)\,dt}$.

Należy zauważyć, że zgodnie z twierdzeniem 3.18, w funkcji $f_X(.)$ znowu nie musimy brać pod uwagę zmiennej normującej.

W twierdzeniu jest mowa o rozkładzie równomiernym na pewnym zbiorze. Intuicyjnie jest to zupełnie zrozumiałym, jeśli pomyślimy o „w pełni losowym, równomiernym rzucie punktem" na ów zbiór. Ściślej mówi o tym następująca definicja.

Definicja 3.19. *Powiemy, że punkt losowy X ma rozkład równomierny na zbiorze $\mathcal{A} \subset \mathbb{R}^p$, jeśli dla dowolnego podzbioru $\mathcal{B}$ tego zbioru $\mathcal{A}$ zachodzi*

$$\mathrm{P}(X \in \mathcal{B}) = \frac{l_p(\mathcal{B})}{l_p(\mathcal{A})} \ , \tag{3.33}$$

gdzie $l_p(.)$ jest p-wymiarową miarą danego zbioru.

Dowód tw. 3.18. Jeśli punkt (U, V) jest punktem losowym o rozkładzie równomiernym na zbiorze $\mathcal{C}_f$, to jego gęstość wyraża się wzorem

$$f_{(U,V)}(u, v) = \frac{1}{l_2(\mathcal{C}_f)} \mathbb{1}_{\mathcal{C}_f}(u, v) \ . \tag{3.34}$$

Wykorzystajmy wzajemnie jednoznaczne przekształcenie

$$X = \frac{V}{U} \ , \ Y = U \ . \tag{3.35}$$

Jakobian tego przekształcenia jest równy

$$\left| \det \begin{pmatrix} 0 & 1 \\ y & x \end{pmatrix} \right| = y \ , \tag{3.36}$$

stąd w nowych zmiennych

$$f_{(X,Y)}(x, y) = \frac{1}{l_2(\mathcal{C}_f)} y \mathbb{1}_{\mathcal{C}_f}(y, xy) = \frac{y}{l_2(\mathcal{C}_f)} \mathbb{1}_{[0;\sqrt{f(x)}]}(y) \ , \tag{3.37}$$

co po odpowiednim całkowaniu prowadzi do gęstości brzegowej zmiennej X

$$f_X(x) = \int f_{(X,Y)}(x, y)\,dy = \frac{1}{l_2(\mathcal{C}_f)} \int_0^{\sqrt{f(x)}} y\,dy = \frac{1}{2l_2(\mathcal{C}_f)} f(x) \ , \tag{3.38}$$

gdzie $\frac{1}{2l_2(\mathcal{C}_f)}$ jest odpowiednią stałą normującą. □

Algorytm jest zatem następujący: należy wygenerować punkt losowy (U, V) o rozkładzie równomiernym na zbiorze $\mathcal{C}_f$, a potem dokonać przekształcenia $X = \frac{V}{U}$, czyli

Algorytm 3.20.

```
repeat
 {(U,V) = GenerujCf
 X = V / U}
until U^2 <= f(X)
return X
```

gdzie funkcja `GenerujCf` służy do generowania punktu z jednostajnego rozkładu na zbiorze $\mathcal{C}_f$. W najprostszym przypadku zbiorem $\mathcal{C}_f$ może być np. prostokąt.

Możemy zauważyć następujące własności zbioru $\mathcal{C}_f$:

1. z definicji $\mathcal{C}_f$ mamy $u \geqslant 0$ (prawa połówka układu współrzędnych)
2. jeśli $f_X(.)$ jest symetryczna względem zera, to zbiór $\mathcal{C}_f$ jest symetryczny względem osi u
3. jeśli $f_X(.)$ jest zdefiniowana tylko dla nieujemnych argumentów, to zbiór $\mathcal{C}_f$ będzie się znajdować w prawej górnej ćwiartce układu u, v, gdyż zachodzi

$$f_X\left(\frac{v}{u}\right) \geqslant 0 \Rightarrow \frac{v}{u} \geqslant 0 \Rightarrow v \geqslant 0 \ .$$

Stosując podstawienie

$$z = \frac{v}{u} \tag{3.39}$$

możemy w parametryczny sposób przedstawić ograniczenie zbioru $\mathcal{C}_f$. Ze (3.39) i ograniczenia zbioru w twierdzeniu 3.18 mamy bowiem

$$u = \sqrt{f(z)} \ , \ v = z\sqrt{f(z)} \ , \tag{3.40}$$

stąd

$$0 \leqslant u \leqslant \sup_z \sqrt{f(z)} \ , \ \inf_z z\sqrt{f(z)} \leqslant v \leqslant \sup_z z\sqrt{f(z)} \ , \tag{3.41}$$

co daje warunki na „zawarcie" zbioru $\mathcal{C}_f$ w pewnym prostokącie.

Przykład 3.21. *Rozpatrzmy zmienną o rozkładzie wykładniczym. Mamy zatem* $f_X(t) = \exp(-\lambda t)$. *Stąd*

$$\sup_t \sqrt{e^{-\lambda t}} = 1 \ , \ \inf_t t\sqrt{e^{-\lambda t}} = 0 \ , \ \sup_t t\sqrt{e^{-\lambda t}} = \frac{2}{\lambda e} \ , \tag{3.42}$$

zatem

$$0 \leqslant u \leqslant 1 \ , \ 0 \leqslant v \leqslant \frac{2}{\lambda e} \tag{3.43}$$

co umożliwia znalezienie prostokąta zawierającego zbiór $\mathcal{C}_f$ *postaci* $[0;1] \times [0; \frac{2}{\lambda e}]$, *a gęstość (dwuwymiarowa), z której generujemy punkty równomiernie z tego prostokąta wyraża się wzorem*

$$g_{\mathcal{C}_f}(t) = \frac{\lambda e}{2} \ . \tag{3.44}$$

Przykład 3.22. *Przyjrzyjmy się rozkładowi normalnemu standardowemu. Mamy*

$$f_X(t) = e^{-\frac{t^2}{2}} \ , \tag{3.45}$$

zatem

$$\sup_t \sqrt{e^{-\frac{t^2}{2}}} = 1 \ , \ \inf_t t\sqrt{e^{-\frac{t^2}{2}}} = -\sqrt{\frac{2}{e}} \ , \ \sup_t t\sqrt{e^{-\frac{t^2}{2}}} = \sqrt{\frac{2}{e}} \ , \tag{3.46}$$

stąd

$$0 \leqslant u \leqslant 1 \ , \ -\sqrt{\frac{2}{e}} \leqslant v \leqslant \sqrt{\frac{2}{e}} \ . \tag{3.47}$$

3.5 Metoda superpozycji rozkładów

Metoda superpozycji (kompozycji) rozkładu polega na przedstawieniu rozważanej gęstości $f_X(.)$ w postaci

$$f_X(t) = \sum_{i=1}^{\infty} p_i f_i(t) \ , \tag{3.48}$$

gdzie $p_i > 0, \sum_{i=1}^{\infty} p_i = 1$, a $f_i(.)$ są gęstościami pewnych rozkładów prawdopodobieństwa. Wzór (3.48) należy rozpatrywać w następujący sposób: w pierwszym kroku generujemy zmienną losową K określoną przez rozkład dyskretny $p_1, p_2, \ldots$, która wskazuje z której gęstości $f_K(.)$ należy wygenerować poszukiwaną zmienną w drugim kroku. Prowadzi to do następującego algorytmu

Algorytm 3.23.

```
K = GenerujK
X = GenerujF (K)
return X
```

przy czym funkcja `GenerujK` generuje zmienną z rozkładu dyskretnego zadanego prawdopodobieństwami $p_1, p_2, \ldots$, a funkcja `GenerujF (K)` to rodzina funkcji generujących zmienną z rozkładu określonego gęstością $f_K(.)$.

W praktyce zazwyczaj wzór (3.48) upraszcza się do postaci ze skończonym rozkładem dyskretnym

$$f_X(t) = \sum_{i=1}^{n} p_i f_i(t) \ . \tag{3.49}$$

W takim przypadku cały przedział określoności $f_X(.)$ dzieli się na rozłączne podzbiory $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$, tak, aby generowanie z poszczególnych gęstości $f_i(.)$ było jak najłatwiejsze. Podejście takie wykorzystaliśmy w przykładzie 3.11, gdzie cała prosta $\mathbb{R}$ została podzielona na dwie części (z symetrii standardowego rozkładu normalnego mieliśmy $p_1 = p_2 = 0,5$), a dla każdej części generowaliśmy zmienną będącą ogonem rozkładu $N(0;1)$ wykorzystując metodę eliminacji z rozkładu wykładniczego.

Zauważmy, że dla (3.49) mamy

$$p_i = \int_{\mathcal{A}_i} f_X(t)\ dt \ , \ f_i(t) = \frac{f_X(t)}{p_i} \mathbb{1}_{\mathcal{A}_i}(t) \ , \tag{3.50}$$

niezbędna jest zatem wiedza na temat wartości prawdopodobieństw, jakie określa gęstość $f_X(.)$ dla poszczególnych zbiorów (przedziałów) $\mathcal{A}_i$.

Przykład 3.24. *Rozpatrzmy rozkład wykładniczy z parametrem $\lambda = 1$. Przedział określoności podzielimy na podzbiory o końcach w kolejnych liczbach całkowitych. Mamy zatem dla zmiennej X z rozkładu wykładniczego*

$$p_i = \mathrm{P}(i-1 \leqslant X < i) = \int_{i-1}^{i} e^{-t} dt = e^{-(i-1)} - e^{-i} = e^{-(i-1)}(1 - e^{-1}) \tag{3.51}$$

oraz

$$f_i = \frac{1}{p_i} e^{-t} \mathbb{1}(i-1 \leqslant t < i) = \frac{e^{-(t-(i-1))}}{1 - e^{-1}} \mathbb{1}(i-1 \leqslant t < i) \ . \tag{3.52}$$

Zastosowanie metody superpozycji w postaci (3.49) jest szczególnie przydatne w połączeniu z metodą eliminacji. Na mniejszych przedziałach jest bowiem często łatwiej znaleźć odpowiednie gęstości $g_i(.)$ dominujące poszczególne „fragmenty" $f_i(.)$.

Wzór (3.48) możemy w ogólniejszy sposób zapisać jako całkę

$$f_X(t) = \int_{\mathcal{Z}} f_z(t)h(z)\,dz\ , \tag{3.53}$$

gdzie $f_z(.)$ dla każdej wartości parametru z jest pewną gęstością, a $h(z)$ jest gęstością pewnego rozkładu określonego na zbiorze $\mathcal{Z}$ (np. prostej rzeczywistej). Wzór ten ma ważną interpretację związaną z gęstościami warunkowymi. Gęstość $f_z(.)$ możemy bowiem interpretować jako gęstość warunkową $f(.|z)$, a $h(z)$ jako gęstość określającą rozkład parametru z. Wtedy $f_X(t)$ jest gęstością bezwarunkową, powstałą poprzez „wycałkowanie" parametru z. Interpretacja ta bliska jest próbnikowi Gibbsa i jednocześnie podpowiada, w jaki sposób możliwe jest otrzymywanie gęstości bezwarunkowej, jeśli dysponujemy gęstością warunkową.

Odpowiedni algorytm dla (3.53) ma postać

Algorytm 3.25.

```
Z = GenerujH
X = GenerujF (Z)
return X
```

najpierw następuje więc wygenerowanie parametru z z gęstości $h(.)$, a w drugim kroku do generowania wykorzystywana jest odpowiednia funkcja gęstości $f_z(.)$.

3.6 Metody generowania z rozkładów dyskretnych

Wśród rozkładów prawdopodobieństwa szczególne miejsce zajmują rozkłady dyskretne, czyli o przeliczalnej liczbie możliwych wartości zmiennej losowej. Najprostszym algorytmem jest metoda bazująca bezpośrednio na metodzie odwracania dystrybuanty

Algorytm 3.26.

```
S = 0
U = GenerujU
I = 0
while S <= U do
 {I = I + 1
 S = S + p_I}
X = I
return X
```

W algorytmie tym generujemy zmienną X o rozkładzie zadanym prawdopodobieństwami $\mathrm{P}(X = 1) = p_1, \mathrm{P}(X = 2) = p_2, \ldots$. Jak łatwo zauważyć, procedura dzieli przedział $[0; 1]$ na odcinki o długościach odpowiadających owym prawdopodobieństwom $p_1, p_2, \ldots$. Następnie algorytm, po wylosowaniu jednostajnie

wartości U z przedziału jednostkowego, „szuka numeru" I odcinka p_I, w którym U się znalazło. Znaleziona wartość jest zwracana jako wynik X.

Powyższy algorytm jest bardzo prosty, jednak może być mało efektywny i zajmować dużo czasu obliczeniowego, jeśli rozkład ma nieskończenie wiele wartości lub prawdopodobieństwa p_i poszczególnych stanów są niezbyt duże. W tego typu przypadkach pomóc może odpowiednie przenumerowanie wartości zmiennej losowej, tak, aby stany bardziej prawdopodobne były przed stanami mniej prawdopodobnymi.

Bardziej zaawansowanym algorytmem jest algorytm ALIAS, który może być stosowany do dyskretnych rozkładów o skończonej liczbie wartości. Załóżmy, że oprócz prawdopodobieństw

$$\mathrm{P}(X=1)=p_1, \mathrm{P}(X=2)=p_2, \ldots, \mathrm{P}(X=m)=p_m \ , \tag{3.54}$$

dysponujemy ciągiem $q_1, q_2, \ldots, q_m$, takim, że $0 \leqslant q_i \leqslant 1$ dla $i=1,2,\ldots,m$ oraz ciągiem $A(1), A(2), \ldots, A(m)$ o wartościach w zbiorze $\{1,2,\ldots,m\}$. Oba te ciągi spełniają przy tym warunek

$$p_i = \left(q_i + \sum_{j:A(j)=i} (1-q_j) \right) \frac{1}{m} \tag{3.55}$$

dla $i=1,2,\ldots,m$. Algorytm ma wtedy następującą postać

Algorytm 3.27.

```
I = GenerujU[m]
U = GenerujU
if U < q_I then X = I
 else X = A (I)
return X
```

gdzie funkcja `GenerujU[m]` generuje wartość z rozkładu jednostajnego na zbiorze $\{1,2,\ldots,m\}$.

Metodę ALIAS najłatwiej zrozumieć na prostym przykładzie.

Przykład 3.28. *Niech dany będzie rozkład prawdopodobieństwa*

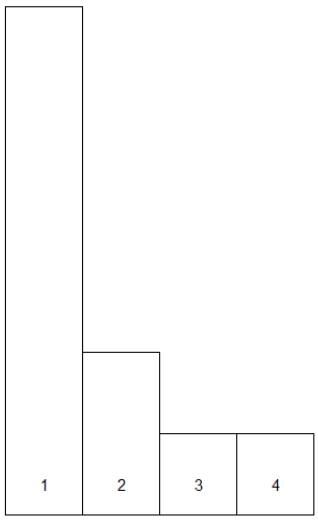
$$p_1=0,6, p_2=0,2, p_3=0,1, p_4=0,1 \ . \tag{3.56}$$

Rozwiązanie: W takim przypadku $m=4$. Prawdopodobieństwa (3.56) mnożymy przez m, otrzymując

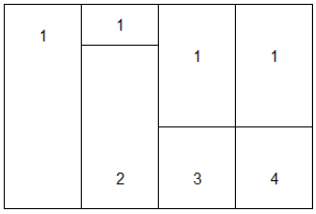
$$mp_1=2,4, mp_2=0,8, mp_3=0,4, mp_4=0,4 \ . \tag{3.57}$$

Wartości te odpowiadają zatem czterem „słupkom jednostkowym", które musimy odpowiednio podzielić (patrz rys. 3.1).

„Pierwszy słupek" ma wysokość 2,4, a powinien mieć jedną jednostkę. „Nadmiarową wysokość" równą 1,4 przenosimy zatem do pozostałych słupków w ten sposób, aby „dopełnić" je do wysokości jeden. W wyniku takiej operacji otrzymujemy odpowiedni podział (patrz rys. 3.2).



Rysunek 3.1: ALIAS – przed podziałem



Rysunek 3.2: ALIAS – po podziale

Podział ten jednocześnie ustala odpowiednie wartości dla ciągów q_i oraz $A(i)$. Mamy zatem

$$q_1 = 1, q_2 = 0,8, q_3 = 0,4, q_4 = 0,4, A(1) = 1, A(2) = 1, A(3) = 1, A(4) = 1 \ . \tag{3.58}$$

Jak widzimy, wartości q_i odpowiadają punktom „przecięcia" na każdym ze „słupków". Natomiast $A(i)$ wskazuje na wartość, która jest „dopełnieniem" do danego „słupka".

Po ustaleniu odpowiednich wartości, algorytm działa w następujący sposób. W pierwszym kroku losowana jest wartość I, która wskazuje na odpowiedni „słupek" z m możliwych. W drugim kroku losowany jest punkt U z danego „słupka". Jeśli U ma wartość mniejszą niż punkt „przecięcia" q_I, to zwracany jest numer „słupka". W przeciwnym przypadku zwracana jest wartość „dopełnienia", czyli $A(I)$. ◇

3.7 Metody szczegółowe

Jak zostało wcześniej wspomniane, oprócz algorytmów ogólnych, które można zastosować do rozmaitych rozkładów, istnieją też metody bardziej szczególne, dzięki którym można uzyskać jeden, szczególny rozkład. Omówimy teraz niektóre z takich metod.

W przeglądzie algorytmów skupimy się na metodach wykorzystywanych przy rozkładzie normalnym, jednym z najczęściej stosowanych rozkładów w statystyce. Wcześniej przedstawiono już metodę odwracania dystrybuanty, metodę eliminacji i metodę ROU dla rozkładu $N(0;1)$. Przypomnijmy sobie, że jeśli zmienna X pochodzi z rozkładu normalnego standardowego, to zmienna $Y = \sigma X + \mu$ ma już dowolny rozkład $N(\mu;\sigma^2)$. Możemy się zatem skupić tylko na zagadnieniu uzyskiwaniu zmiennej z rozkładu $N(0;1)$.

Algorytm prymitywny Jednym z najprostszych generatorów rozkładu normalnego ma postać

Algorytm 3.29.

```
X = 0
for I = 1 to 12 do
 X = X + GenerujU
X = X - 6
return X
```

Algorytm ten odpowiada obliczeniu

$$X = \sum_{i=1}^{12} U_i - 6 \ , \tag{3.59}$$

gdzie $U_1, U_2, \ldots, U_{12} \overset{iid}{\sim} U$. Jak łatwo zauważyć, $\mathbb{E}\, X = 0, \mathrm{Var}\, X = 1$. Jednocześnie zmienna X ma rozkład zbliżony do normalnego, dzięki zachodzeniu centralnego twierdzenia granicznego (patrz twierdzenie 1.33).

Algorytm Boxa-Mullera Metoda ta pozwala na wygenerowanie *dwóch* zmiennych niezależnych X_1, X_2 z rozkładu $N(0;1)$.

Algorytm 3.30.

```
U1 = GenerujU
U2 = GenerujU
Phi = 2 * Pi * U1
R = Sqrt ( - 2 * Ln ( U2 ))
X1 = R * Cos ( Phi )
X2 = R * Sin ( Phi )
return X1, X2
```

Jak widzimy, metoda ta wymaga wygenerowania dwóch zmiennych niezależnych z rozkładu jednostajnego na przedziale jednostkowym U_1, U_2. Zmienna U_1 jest następnie mnożona przez 2π, dając współrzędną kątową ϕ. Zmienna U_2 jest z kolei przekształcana do zmiennej z rozkładu wykładniczego $\mathrm{Exp}(1/2)$, dając długość promienia wodzącego R w biegunowym układzie współrzędnych. Wynikowe zmienne, powstałe z przekształcenia ϕ i R okazują się niezależnymi zmiennymi z rozkładu $N(0;1)$.

Z praktycznego punktu widzenia jest to bardzo interesująca metoda, dość szybka numerycznie (jeśli tylko dysponujemy odpowiednio szybkimi procedurami liczącymi logarytmy i funkcje trygonometryczne). Zaletą jest również generowanie od razu dwóch zmiennych – jeśli jednak potrzebujemy tylko pojedynczej zmiennej, staje się to wadą tej metody.

Algorym Marsaglii Ten algorytm jest w wielu punktach podobny do poprzedniego.

Algorytm 3.31.

```
repeat
 {U1 = GenerujU;
U2 = GenerujU;
U1 = 2 * U1 - 1;
U2 = 2 * U2 - 1;
W = U1^2 + U2^2}
until W < 1
C = Sqrt ( - 2 * W^(-1) * Ln ( W ) )
X1 = C * U1
X2 = C * U2
return X1, X2
```

W metodzie tej generujemy dwie zmienne U_1, U_2 z rozkładu jednostajnego na przedziale jednostkowym. Następnie przekształcamy te zmienne, aby pochodziły z rozkładu jednostajnego na przedziale $[-1;1]$, czyli zawierały się w kwadracie $[-1;1] \times [-1;1]$. Punkt z takiego kwadratu jest generowany dopóty, dopóki nie będzie się zawierał w okręgu jednostkowym. Sprawdzeniu tego warunku służy obliczanie zmiennej W, czyli kwadratu odległości punktu od środka układu współrzędnych. Jeśli punkt zawiera się we wspomnianym okręgu, zostają obliczone nowe zmienne X_1, X_2. Zauważmy, że niezależne współrzędne U_1, U_2 wskazują „kąt" dla nowych zmiennych X_1, X_2, zaś zmienna W dokonuje przesunięcia na prostej wyznaczonej przez (U_1, U_2). Ów kąt ma rozkład jednostajny, zaś kwadrat odległości R^2 dla nowego punktu możemy wyznaczyć z zależności

$$R^2 = C^2\left(U_1^2 + U_2^2\right) = C^2 W = \left(-\frac{2}{W}\ln W\right) W = -2\ln W\ , \tag{3.60}$$

co wskazuje na bezpośredni związek z algorytmem Boxa-Mullera.

Algorytm Marsaglii-Braya Ze względu na dużą liczbę stałych i parametrów występujących w tej metodzie, opiszemy ją trochę bardziej ogólnie. Jest ona jednak na tyle ciekawa i dobrze ilustruje metodę superpozycji (patrz rozdział 3.5), że warto o niej wspomnieć.

Rozkład $N(0;1)$ w tej metodzie otrzymujemy dekomponując jego poszczególne składowe. Ogony tego rozkładu możemy wygenerować korzystając z metod zaprezentowanych na ćwiczeniach (np. metodą eliminacji z uogólnionego rozkładu wykładniczego). Przypuśćmy, że obcięcia ogonów dokonujemy dla wartości 3, tzn. metody dla ogonów zostaną użyte na przedziałach $(-\infty;-3]\cup[3;\infty)$. Prawdopodobieństwo, że zmienna z rozkładu normalnego pochodzić będzie z sumy tych przedziałów wynosi

$$p_4 = \mathrm{P}(|X| > 3) \approx 0,0027 \ . \tag{3.61}$$

Pozostaje zatem rozpatrzeć przedział $[-3;3]$. Jak się okazuje, gęstość rozkładu normalnego standardowego na tym przedziale przybliżyć możemy sklejeniem paraboli

$$f_1(t) = \begin{cases} \frac{3-t^2}{8} & \text{dla } |t| < 1 \\ \frac{(3-|t|)^2}{16} & \text{dla } 1 \leqslant |t| \leqslant 3 \\ 0 & \text{poza tym} \end{cases} \ . \tag{3.62}$$

Dla tego przybliżenia dobieramy maksymalne p_1 spełniające warunek

$$f_{N(0;1)}(t) - p_1 f_1(t) \geqslant 0 \ , \tag{3.63}$$

co dla $|t| \leqslant 3$ daje $p_1 = \frac{16}{\sqrt{2\pi e}} \approx 0,86$. Z takim więc, stosunkowo dużym prawdopodobieństwem będziemy generować zmienną z gęstości $f_1(.)$, która jest jednocześnie sumą $U_1 + U_2 + U_3$ niezależnych zmiennych losowych o rozkładzie jednostajnym na przedziale jednostkowym.

„Resztę", czyli funkcję $f_{N(0;1)}(.) - p_1 f_1(.)$ należy poddać dalszej dekompozycji. W tym celu wykorzystujemy gęstość

$$f_2(t) = \begin{cases} \frac{4}{9}\left(\frac{3}{2} - |t|\right) & \text{dla } |t| \leqslant \frac{3}{2} \\ 0 & \text{poza tym} \end{cases} \ . \tag{3.64}$$

Gęstość ta jest gęstością rozkładu zmiennej losowej

$$\frac{3(U_1 + U_2 - 1)}{2} \ , \tag{3.65}$$

gdzie U_1, U_2 są niezależnymi zmiennymi losowymi z rozkładu jednostajnego na przedziale jednostkowym. Podobnie jak poprzednio, wyznaczamy taką maksymalną wartość p_2 dla której

$$f_{N(0;1)}(t) - p_1 f_1(t) - p_2 f_2(t) \geqslant 0 \tag{3.66}$$

na przedziale $[-3;3]$. Otrzymujemy stąd $p_2 \approx 0,111$, co oznacza, że z prawdodopodobieństwem $p_3 = 1 - p_1 - p_2 - p_4 \approx 0,0226$ trzeba losować z innego rozkładu, o gęstości „resztowej", tzn.

$$f_3(t) = \frac{1}{p_3}\left(f_{N(0;1)}(t) - p_1 f_1(t) - p_2 f_2(t) - p_4 f_4(t)\right) \ . \tag{3.67}$$

Gęstość tą, która jest liniową mieszaniną funkcji wykładniczej i wielomianu, można oszacować z góry na przedziale $[-3;3]$. Oznacza to zatem, że możemy ją uzyskać korzystając bezpośrednio z metody eliminacji.

Jeśli chodzi o inne rozkłady, to istnieje cały szereg odpowiednich algorytmów. Jako przykład podamy następujące twierdzenie, pozwalające generować z rozkładu beta.

Twierdzenie 3.32. *Jeśli zmienne $U, V \sim U[0;1]$ i są niezależne, to warunkowy rozkład zmiennej*

$$X = \frac{U^{\frac{1}{\alpha}}}{U^{\frac{1}{\alpha}} + V^{\frac{1}{\beta}}} \tag{3.68}$$

pod warunkiem

$$U^{\frac{1}{\alpha}} + V^{\frac{1}{\beta}} \leqslant 1 \tag{3.69}$$

jest rozkładem beta o parametrach (α, β) określonym funkcją gęstości

$$f_X(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1}(1-t)^{\beta-1} \tag{3.70}$$

dla $t \in (0;1)$.

Z twierdzenia tego bezpośrednio wynika sposób generowania zmiennej X. Losujemy mianowicie dwie zmienne niezależne U, V z rozkładu jednostajnego na przedziale jednostkowym i jeśli spełniają one warunek (3.69), zmienna X dana wzorem (3.68) będzie zmienną z rozkładu Beta.

3.8 Wielowymiarowe zmienne losowe

Przejdziemy teraz do omawiania metod generowania wielowymiarowych zmiennych losowych. Odpowiedni wektor losowy o p składowych $\left(X^{(1)}, X^{(2)}, \ldots X^{(p)}\right)$ oznaczać będziemy przez $\mathbb{X}$. Zauważmy, że zgodnie z oznaczeniami z rozdziału 1.4.10, indeks górny będzie teraz grać rolę wskaźnika dla współrzędnej, a indeks dolny – jak dotychczas – numeru zmiennej. W ten sposób

$$\mathbb{X}_i = \left(X_i^{(1)}, X_i^{(2)}, \ldots X_i^{(p)}\right) \tag{3.71}$$

oznaczać będzie i-tą zmienną w ciągu o p składowych.

Teoretycznie rzecz biorąc, jeśli poszczególne składowe są niezależnymi zmiennymi losowymi o tych samych rozkładach, to w celu wygenerowania losowego ciągu $\mathbb{X}_1, \mathbb{X}_2, \mathbb{X}_3, \ldots$ wystarczy stworzyć za każdym razem p niezależnych zmiennych losowych korzystając np. z poznanych wcześniej metod. Zauważmy zresztą, że niektóre metody (np. algorytm Boxa-Mullera) „z samej swej natury" tworzą zmienne wielowymiarowe. W przypadku tego algorytmu z dwóch zmiennych U_1, U_2 otrzymywaliśmy dwie niezależne zmienne X_1, X_2 o rozkładzie normalnym standardowym. Wystarczy zatem określić

$$\mathbb{X}_1 = (X_1, X_2)\ ,\ \ldots \tag{3.72}$$

i wielokrotnie korzystając z algorytmu Boxa-Mullera w „naturalny" sposób otrzymujemy dwuwymiarową zmienną losową o niezależnych współrzędnych będących rozkładami normalnymi standardowymi. Rozkład taki zapisywać będziemy zgodnie z notacją z rozdziału 1.4.10

$$\mathbb{X}_1, \mathbb{X}_2, \ldots \overset{iid}{\sim} N(0, 1 * \mathbb{I}) \tag{3.73}$$

Przykład 3.33. *W metodzie ROU niezbędne było generowanie zmiennej o rozkładzie równomiernym na pewnym zbiorze $\mathcal{C}_f$. Był to rozkład dwuwymiarowy o gęstości zadanej np. przez funkcję*

$$g_{\mathcal{C}_f} = \frac{\lambda e}{2} \tag{3.74}$$

(patrz przykład 3.21). Jest to gęstość dla rozkładu dwuwymiarowego na prostokącie $[0;1] \times [0; \frac{2}{\lambda e}]$ o dwóch niezależnych współrzędnych. W celu uzyskania odpowiedniego wektora losowego wystarczy więc wygenerować jedną zmienną z rozkładu jednostajnego na przedziale jednostkowym i drugą zmienną (niezależnie!) z rozkładu jednostajnego na przedziale $[0; \frac{2}{\lambda e}]$. Odpowiedni algorytm ma więc postać

Algorytm 3.34.

```
U1 = GenerujU
U2 = GenerujU
U2 = 2 * U2 / (lambda * e)
X = ( U1, U2 )
return X
```

Ogólniej rzecz biorąc, do generowania ciągów wielowymiarowych zmiennych losowych (również jeśli ich poszczególne komponenty nie są od siebie wzajemnie niezależne) zastosować możemy poznane wcześniej metody: metodę eliminacji i metodę superpozycji.

Jeśli chodzi o metodę eliminacji, problemem może być znalezienie odpowiedniej gęstości dominującej. Co więcej, proste oszacowania mogą być bardzo mylące. Wynika to z problemu zwanego *przekleństwem wielowymiarowości* (lub *demonem wielowymiarowości* – w j. ang. *curse of dimensionality*). Jak bowiem pamiętamy, w metodzie eliminacji część punktów ulega *eliminacji* – wraz ze wzrostem wymiaru może to prowadzić do odrzucania *coraz większej* liczby punktów.

Przykład 3.35. *Rozpatrzmy metodę eliminacji dla generowania punktów o rozkładzie równomiernym z p wymiarowej kuli jednostkowej B^p. Zauważmy, że wystarczy wygenerować punkt $\mathbb{X}$ z jednostajnego rozkładu na p wymiarowej kostce U^p, tzn.*

$$\mathbb{X} = (U_1, U_2, \ldots, U_p) \ , \tag{3.75}$$

gdzie $U_1, \ldots, U_p$ są zmiennymi iid *z $U[-1;1]$. Jeśli tylko punkt $\mathbb{X}$ znajduje się we wnętrzu kuli B^p, tzn. jego współrzędne spełniają warunek*

$$U_1^2 + U_2^2 + \ldots + U_p^2 \leqslant 1 \ , \tag{3.76}$$

zostaje on zaakceptowany. W tym przykładzie gęstością dominującą jest zatem rozkład jednostajny na U^p, a algorytm ma postać

Algorytm 3.36.

```
repeat
 {for i=1 to p do
   U[i] = 2 * GenerujU -1}
until U[1]^2 + U[2]^2 + ... + U[p]^2 <=1
X = ( U[1], U[2],..., U[p] )
return X
```

Prawdopodobieństwo $\mathrm{P}(p)$ akceptacji punktu w powyższym algorytmie jest równe stosunkowi objętości kuli

$$|B^p| = \frac{2\pi^{\frac{p}{2}}}{p\Gamma\frac{p}{2}} \tag{3.77}$$

do objętości kostki U^p, czyli

$$\mathrm{P}(p) = \frac{|B^p|}{|U^p|} = \frac{\pi^{\frac{p}{2}}}{p2^{p-1}\Gamma\frac{p}{2}} \ , \tag{3.78}$$

przy czym

$$\lim_{p\to+\infty} \frac{|B^p|}{|U^p|} = 0 \ . \tag{3.79}$$

Oczekiwana liczba wylosowanych punktów zanim nastąpi akceptacja pierwszego z nich $\mathbb{E}\, N_p$, jest równa odwrotności prawdopodobieństwa (3.78). Sprawia to, że efektywność przedstawionego algorytmu jest bardzo niska. Przykładowo mamy

p	$\mathrm{P}(p)$	$\mathbb{E}\, N_p$
2	$7,8 \cdot 10^{-1}$	$1,27$
5	$1,6 \cdot 10^{-1}$	$6,08$
10	$2,5 \cdot 10^{-3}$	$4 \cdot 10^2$
20	$2,5 \cdot 10^{-8}$	$4 \cdot 10^7$
50	$1,5 \cdot 10^{-28}$	$6,5 \cdot 10^{27}$

Jak widzimy, już nawet przy pięciu wymiarach, potrzeba średnio sześciu przebiegów algorytmu, zanim odpowiedni punkt zostanie wygenerowany.

Dla przypadku $p = 2$ lepszym rozwiązaniem może być algorytm bazujący na współrzędnych biegunowych

Algorytm 3.37.

```
Phi = 2 * Pi * GenerujU
X1 = cos (Phi)
X2 = sin (Phi)
u = GenerujU
Y1 = Sqrt (U) * X1
Y2 = Sqrt (U) * X2
return ( X1,X2 )
```

W algorytmie tym najpierw generujemy punkt położony równomiernie na okręgu o współrzędnych (X_1, X_2). Następnie skalujemy go, otrzymując punkt położony równomiernie w kole o współrzędnych (Y_1, Y_2).

Niestety, uogólnienie tego algorytmu na wyższe wymiary dla współrzędnych sferycznych nie daje oczekiwanych rezultatów.

Przykład 3.38. *Rozpatrz uogólnienie algorytmu generowania punktu o rozkładzie równomiernym na sferze dla $p = 3$ bazującego na współrzędnych sferycznych. Dlaczego ten algorytm nie działa? Podpowiedź: zwróć uwagę na pola i otrzymywane algorytmem prawdopodobieństwa dla strefy wokół bieguna i pasa dookoła równika.*

Dlatego znacznie częściej stosuje się algorytm bazujący na odpowiednim unormowaniu zmiennych. Jeśli rozpatrujemy rozkład jednostajny na kuli, to odpowiedni algorytm ma postać

Algorytm 3.39.

```
for i = 1 to p do
 Z[i] = GenerujNormalnyStd
R^2 = Z[1]^2 + ... + Z[p]^2
for i = 1 to p do
 Y[i] = Z[i] / R
U = GenerujU
R1 = U^(1/p)
for i = 1 to p do
 X[i] = Y[i] * R1
return ( X[1], ..., X[p] )
```

Algorytm ten generuje p zmiennych losowych Z_i, każda z niezależnego standardowego rozkładu normalnego. Są one następnie normowane względem odległości R w celu otrzymania zmiennej $(Y_1, \ldots, Y_p)$, która pochodzi z rozkładu równomiernego na sferze jednostkowej. Następnie losowana jest dodatkowa zmienna z rozkładu jednostajnego na przedziale jednostkowym. Po jej przekształceniu, wcześniej uzyskany punkt ze sfery jest przesuwany do nowej pozycji $(X_1, \ldots, X_p)$ tak, aby uzyskać rozkład jednostajny na kuli.

Innym podejściem, wykorzystującym wielowymiarowość problemu, może być przedstawienie docelowej gęstości jako odpowiedniego iloczynu gęstości warunkowych

$$f_{\mathbb{X}}\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}\right) = \\ = f_1\left(x^{(1)}\right) f_2\left(x^{(2)}|x^{(1)}\right) \ldots f_p\left(x^{(p)}|x^{(1)}, \ldots, x^{(p-1)}\right) \ . \quad (3.80)$$

Innymi słowy, dokonujemy generowania po „kolejnych współrzędnych". Zaczynamy od „osi" $x^{(1)}$, potem znając już wartość tej zmiennej (czyli warunkowo), generujemy z „osi" $x^{(2)}$, itd. Jest to bardzo „eleganckie" podejście, wymaga jednak przedstawienia szukanej gęstości $f_{\mathbb{X}}(.)$ w postaci (3.80), co nie zawsze jest łatwe lub nawet możliwe. Jeśli jednak jesteśmy w stanie przedstawić gęstość jako odpowiedni iloczyn, to algorytm ma wtedy postać

Algorytm 3.40.

```
X[1] = Generujf[1]
for i = 2 to p do
 X[i] = Generujf[i] ( X[1], X[2], ..., X[i-1] )
return ( X[1], X[2], ..., X[p] )
```

3.8.1 Wielowymiarowy rozkład normalny

Podobnie jak poprzednio, skupimy się przede wszystkim na rozkładzie normalnym, jako na najpowszechniej występującym i najczęściej stosowanym w statystyce. Jeśli chcemy wygenerować zmienne losowe

$$\mathbb{X}_1, \mathbb{X}_2, \ldots \overset{iid}{\sim} N(\mu, \sigma^2 \cdot \mathbb{I}) \ , \quad (3.81)$$

czyli o niezależnych poszczególnych składowych (tzn. $\mathrm{Cov}\left(X^{(i)}, X^{(j)}\right) = 0$ dla $i \neq j$), to jest to możliwe przy wykorzystaniu wcześniej poznanych metod dla jednowymiarowych zmiennych losowych. Jeśli jednak macierz kowariancji

$$\mathrm{VAR}\,\mathbb{X} = \mathbb{W} = \begin{pmatrix} \mathrm{Var}\, X^{(1)} & \mathrm{Cov}(X^{(1)}, X^{(2)}) & \dots & \mathrm{Cov}(X^{(1)}, X^{(p)}) \\ \mathrm{Cov}(X^{(2)}, X^{(1)}) & \mathrm{Var}\, X^{(2)} & \dots & \mathrm{Cov}(X^{(2)}, X^{(p)}) \\ \dots & & & \\ \mathrm{Cov}(X^{(p)}, X^{(1)}) & \mathrm{Cov}(X^{(p)}, X^{(2)}) & \dots & \mathrm{Var}\, X^{(k)} \end{pmatrix}, \tag{3.82}$$

ma bardziej skomplikowaną postać, niezbędne jest zastosowanie innych metod.

Dla ułatwienia załóżmy, że $\mathbb{E}\,\mathbb{X} = 0$ i macierz kowariancji $\mathbb{W}$ daje się przedstawić jako iloczyn

$$\mathbb{W} = \mathbb{C}\mathbb{C}^T \tag{3.83}$$

dla pewnej nieosobliwej macierzy $\mathbb{C}$. W takim przypadku, jeśli $\mathbb{Y}$ ma wielowymiarowy rozkład normalny $N(0, 1 \cdot \mathbb{I})$, to po użyciu przekształcenia

$$\mathbb{X} = \mathbb{C}\mathbb{Y} \tag{3.84}$$

zachodzi

$$\mathbb{X} \sim N(0, \mathbb{W}) \,. \tag{3.85}$$

Wzór (3.84) przedstawia więc praktyczny sposób na generowanie odpowiedniej wielowymiarowej zmiennej losowej z rozkładu normalnego. Wystarczy jedynie, korzystając z wcześniej poznanych metod, wygenerować p zmiennych losowych z niezależnych rozkładów normalnych standardowych i skorzystać z liniowego przekształcenia (3.84) w celu otrzymania pożądanego rozkładu normalnego.

Jak więc widzimy, wymaganym krokiem jest skonstruowanie odpowiedniej nieosobliwej macierzy $\mathbb{C}$. Możemy w tym celu wykonać dekompozycję na macierz dolnotrójkątną (czyli z samymi zerami powyżej głównej przekątnej), wykorzystując metodę zwaną dekompozycją Choleskiego. Jest ona zdefiniowana wzorami

$$c_{i,i} = \sqrt{\left(w_{i,i} - \sum_{k=1}^{i-1} c_{i,k}^2 \right)} \tag{3.86}$$

$$c_{j,i} = \frac{w_{j,i} - \sum_{k=1}^{i-1} c_{j,k} c_{i,k}}{c_{i,i}} \,, \tag{3.87}$$

gdzie $w_{i,j}$ i $c_{i,j}$ są odpowiednimi wyrazami w komórkach macierzy $\mathbb{W}$ i $\mathbb{C}$.

Przykład 3.41. *Jeśli macierz $\mathbb{W}$ ma wymiar 2 na 2, to możemy skorzystać z dekompozycji Choleskiego, uzyskując*

$$\mathbb{C} = \begin{pmatrix} \sqrt{w_{1,1}} & 0 \\ \frac{w_{2,1}}{c_{1,1}} & \sqrt{w_{2,2} - c_{2,1}^2} \end{pmatrix} \,. \tag{3.88}$$

Możliwe jest również zastosowanie dekompozycji (3.80). W takim przypadku generowanie z poszczególnych składowych dokonywane jest z odpowiednich warunkowych rozkładów normalnych.

3.8.2 Metoda przekształceń

Zastosowanie liniowego przekształcenia w (3.84) wskazuje na inną możliwość generowania wielowymiarowych zmiennych losowych, jako przekształceń ze znanych już rozkładów zmiennych losowych. Jeśli wielowymiarowa zmienna losowa $\mathbb{Y}$ ma gęstość określoną funkcją $f_{\mathbb{Y}}(.)$ oraz istnieje wzajemnie jednoznaczne przekształcenie $h : \mathbb{R}^p \to \mathbb{R}^p$ klasy C^1 oraz odwrotne do niego przekształcenie h^{-1} też jest klasy C^1, to wtedy zmienna

$$\mathbb{X} = h(\mathbb{Y}) \tag{3.89}$$

ma rozkład określony gęstością

$$g_{\mathbb{X}}(\mathbf{x}) = f_{\mathbb{Y}}(h^{-1}(\mathbf{x})) \left|\det \left(h^{-1}\right)' (\mathbf{x})\right| \ . \tag{3.90}$$

3.8.3 Pojęcie kopuły

Pojęcie kopuły jest coraz powszechniej używane w literaturze statystycznej. Dla ułatwienia, przedstawimy tutaj tylko kopuły dwuwymiarowe, choć istnieją odpowiednie rozszerzenia na przypadki więcej wymiarowe.

Kopułą jest taki wielowymiarowy rozkład łączny określony na kwadracie jednostkowym, dla którego każdy rozkład brzegowy jest rozkładem jednostajnym na przedziale jednostkowym. Dokładniej, kopułą jest funkcja $C : [0;1] \times [0;1] \to [0;1]$ spełniająca warunki:

1. $C(0,t) = C(t,0) = 0$ dla dowolnego $t \in [0;1]$
2. $C(1,t) = C(t,1) = t$ dla dowolnego $t \in [0;1]$
3. $C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geqslant 0$ dla dowolnych $u_1, u_2, v_1, v_2 \in [0;1]$ i $u_1 \geqslant u_2$ i $v_1 \geqslant v_2$

Z pojęciem kopuły jest związanie twierdzenie

Twierdzenie 3.42 (Sklar)**.** *Niech $H(.,.)$ będzie dwuwymiarową dystrybuantą łączną, a $F_x(.)$ i $G_y(.)$ odpowiednimi dystrybuantami brzegowymi. Istnieje wtedy kopuła C spełniająca warunek*

$$H(x,y) = C\left(F_x(x), G_y(y)\right) \ , \tag{3.91}$$

a jeśli $F_x(.)$ i $G_y(.)$ są ciągłe, to przedstawienie to jest jednoznaczne.

Twierdzenie to umożliwia spojrzenie na kopułę jak na narzędzie służące do generowania pewnych rozkładów dwuwymiarowych – podaje bowiem zależność pomiędzy dystrybuantami brzegowymi a dystrybuantą łączną.

Rozdział 4

Generowanie procesów stochastycznych

W niniejszym rozdziale przyjrzymy się pokrótce metodom wykorzystania poznanych wcześniej generatorów liczb (pseudo)losowych do symulowania procesów stochastycznych.

4.1 Proces Poissona

Najprostszą metodą generowania procesu Poissona jest skorzystanie z zależności pomiędzy rozkładem wykładniczym Exp(λ) a procesem Poissona $\mathcal{N}(\lambda)$. Jak wiadomo, odstęp czasu pomiędzy kolejnymi „skokami" w procesie Poissona jest określony rozkładem wykładniczym. Stąd dla ustalonej wartości czasu t

Algorytm 4.1.

```
N = 0
S_0 = 0
while S < t do
 {T = GenerujExp (lambda)
 S = S + T
 N = N + 1}
return N
```

Innymi słowy, generujemy ciąg zmiennych *iid* $T_1, T_2, \ldots$ z rozkładu wykładniczego Exp(λ), aby na ich podstawie utworzyć sumy skumulowane

$$S_1 = T_1, S_2 = T_1 + T_2, S_3 = T_1 + T_2 + T_3, \ldots \ , \tag{4.1}$$

czyli sumy okresów pomiędzy skokami w procesie Poissona. Suma ta jest tworzona aż do momentu przekroczenia określonego wcześniej momentu t i zostaje wtedy zwrócona wartość liczby przyrostów w procesie $\mathcal{N}(\lambda)$.

Jak łatwo zauważyć, algorytm ten ma pewną wadę – dla dużych wartości λ (krótkie okresy pomiędzy skokami) lub dużych wartości t generacja wartości $\mathcal{N}_t$ może zajmować dużo czasu. Dlatego skorzystać można z następującego twierdzenia.

Twierdzenie 4.2. *Przy ustalonej wartości procesu $\mathcal{N}_t = n$, warunkowy rozkład wektora $(S_1, S_2, \ldots, S_n)$ jest taki sam jak rozkład wektora statystyk pozycyjnych $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$, otrzymanych dla rozkładu $U[0;t]$.*

Otrzymujemy dzięki temu następujący algorytm

Algorytm 4.3.

```
N = GenerujPoiss (lambda * t)
for I = 1 to N do
 U_I = GenerujU
(X_{1:N},X_{2:N},...,X_{N:N}) = Uporzadkuj (U_1,U_2,...,U_N)
for I = 1 to N do
 S_I = t * X_{I:N}
```

W pierwszym kroku algorytmu generujemy *od razu* wartość procesu $\mathcal{N}(\lambda)$ w chwili t, czyli N. Wartość ta pochodzi z rozkładu Poissona o parametrze λt, co wynika z definicji procesu Poissona (patrz definicja 1.35). Następnie, przy ustalonej wartości N, generujemy N niezależnych zmiennych z rozkładu jednostajnego na przedziale jednostkowym. Zmienne te porządkujemy, otrzymując statystyki pozycyjne $(X_{1:N}, X_{2:N}, \ldots, X_{N:N})$. Statystyki owe skalujemy poprzez pomnożenie przez t, uzyskując ciąg skumulowanych sum okresów pomiędzy skokami procesu Poissona $(S_1, S_2, \ldots, S_N)$

4.2 Proces Wienera

Generowanie procesu Wienera jest możliwe poprzez bezpośrednie bazowanie na definicji 1.36. Załóżmy, że interesują nas wartości tego procesu w chwilach $t_1 < t_2 < \ldots < t_n$, czyli $\mathcal{W}_{t_1}, \mathcal{W}_{t_2}, \ldots, \mathcal{W}_{t_n}$. Wtedy algorytm przyjmuje następującą postać

Algorytm 4.4.

```
W_0 = 0
for I = 1 to n do
 {N = GenerujN (0, t_I - t_{I-1})
 W_I = W_{I-1} + N}
return (W_1,...,W_n)
```

Algorym ten generuje kolejne przyrosty w procesie Wienera o rozkładach $N(0, t_I - t_{I-1})$. Następnie przyrosty te są dodawane do kolejnych zmiennych

$$\mathcal{W}_{t_1} = N_{t_1}, \mathcal{W}_{t_2} = \mathcal{W}_{t_1} + N_{t_2}, \ldots, \mathcal{W}_{t_n} = \mathcal{W}_{t_{n-1}} + N_{t_n} \ , \tag{4.2}$$

dając wartości procesu Wienera w odpowiednich momentach czasu.

Rozdział 5

Metody Monte Carlo

Metodologia obliczeń, nazwana metodami Monte Carlo, została zaproponowana w latach 40-tych ubiegłego wieku przez dwóch wielkich matematyków – Stanisława Ulama i Nicholasa Metropolisa (patrz [26]). Początkowo metody te zostały wykorzystane w zastosowaniach militarnych przy słynnym projekcie Manhattan, aby w ciągu następnych kilkudziesięciu lat podbić świat matematycznych symulacji komputerowych i znaleźć zastosowanie w wielu praktycznych i teoretycznych problemach.

Sukces odniesiony przez metody Monte Carlo (w skrócie zwane metodami MC) został następnie zdyskontowany przez ich rozwinięcie wykorzystujące teorię łańcuchów Markowa – metody Markov Chain Monte Carlo (skrótowo określane jako MCMC, patrz np. [5]). Metodologia MCMC zaproponowana została w latach 50-tych ubiegłego wieku przez Metropolisa, Rosenbluthów i Tellerów (patrz [25]).

Obecnie lista dziedzin wiedzy, czasami nawet bardzo odlegle związanych z matematyką stosowaną i statystyką, w których wykorzystywane są metody MC i MCMC jest nadzwyczaj imponująca — poczynając od zadań znajdowania całek i zagadnień optymalizacyjnych, poprzez generowanie zmiennych losowych z analitycznie skomplikowanych rozkładów prawdopodobieństwa, estymację złożonych statystyk testowych, zagadnienia z teorii ruiny dla matematyki ubezpieczeniowej, symulacyjne wyznaczanie cen instrumentów finansowych, odszumianie obrazów, problemy związane z wnioskowaniem bayesowskim, aż na zagadnieniach z fizyki, psychologii i medycyny kończąc. Opracowane zostały również specjalne, uniwersalne programy komputerowe wspomagające użycie tych metod w praktyce, jak np. BUGS (patrz np. [33]).

Mimo ogromnego rozwoju metod MC i MCMC, ich stosowanie nierozerwalnie wiąże się nadal z kwestią uciążliwej *diagnostyki zbieżności* rozpatrywanego algorytmu.

Stanisław Ulam w następujący sposób opisał moment wynalezienia metody Monte Carlo (cytat za [7]):

> *The first thoughts and attempts I made to practice [the Monte Carlo Method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully? After spending a lot of time try-*

ing to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays. This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later [in 1946, I] described the idea to John von Neumann, and we began to plan actual calculations.

Jak wynika z powyższego cytatu, największą zasługą Ulama było dostrzeżenie siły zastosowania statystycznych symulacji wykonywanych za pomocą pierwszych komputerów do przybliżonego rozwiązywania praktycznych problemów ilościowych. Należy zwrócić uwagę, że podobne, symulacyjne podejście było znane i wykorzystywane już wcześniej, choć raczej jako swoista ciekawostka na „obrzeżach" statystyki, niż w jej głównym nurcie. W szczególności dopiero Ulam wskazał na możliwość zastosowania komputerów. Klasycznym przykładem jest tutaj tzw. „problem igły Buffona".

5.1 Zagadnienie całkowania metodą MC

W wielu przypadkach, znanych zarówno z teorii, jak i zastosowań, okazuje się, że całka z ustalonej funkcji nie daje się wyrazić bezpośrednio przy pomocy funkcji elementarnych. Zjawisko takie nazywa się często brakiem postaci analitycznej dla całki z funkcji.

W takim przypadku do obliczenia poszukiwanej całki z funkcji konieczne jest wykorzystanie metod numerycznych. Oprócz metod numerycznych bazujących na deterministycznych algorytmach, takich jak np. sumy Riemanna, zasada trapezoidu, zasada Simpsona, funkcje sklejane (splajny) itp., można skorzystać także z metody MC. Polega ona na zapisaniu całki z poszukiwanej funkcji

$$\int_{\mathcal{X}} h^*(x)dx \tag{5.1}$$

jako równoważnej całki z iloczynu dwóch innych, odpowiednio dobranych funkcji $h(x)$ i $f(x)$.

Problem znalezienia poszukiwanej całki sprowadza się wtedy do całkowania iloczynu funkcji o następującej postaci

$$\mathbb{E}_f h(X) = \int_{\mathcal{X}} h(x)f(x)dx \ , \tag{5.2}$$

gdzie $f(x)$ jest gęstością pewnego rozkładu prawdopodobieństwa o nośniku $\mathcal{X}$. Ze względu na praktyczną naturę problemu, zakładać będziemy, iż $\mathcal{X} \subset \mathbb{R}^p$, gdzie p jest pewną liczbą naturalną. Oczywistym warunkiem poprawności równości (5.2) jest istnienie odpowiedniej wartości oczekiwanej. Dlatego w dalszej części pracy zakładać zawsze będziemy, że $\mathbb{E}_f h(x) < \infty$.

Zauważmy, że formuła (5.2) jest bardzo ogólną postacią dla problemu całkowania, gdyż dla dowolnie wybranej funkcji $h(x)$ o zwartym nośniku $\mathcal{X}$, jej całkę

możemy zapisać jako

$$\int_{\mathcal{X}} h(x)dx = \int_{\mathcal{X}} |\mathcal{X}|h(x)\frac{1}{|\mathcal{X}|}dx = |\mathcal{X}|\mathbb{E}_{|\mathcal{X}|}h(X) \ , \tag{5.3}$$

gdzie $|\mathcal{X}|$ jest mocą zbioru $\mathcal{X}$ w przypadku skończoności zbioru $\mathcal{X}$ lub p-wymiarową objętością przestrzeni $\mathcal{X}$ w przypadku, gdy $\mathcal{X} \subset \mathbb{R}^p$. W oczywisty sposób, w (5.3) rolę gęstości $f(x)$ pełni gęstość rozkładu jednostajnego na $\mathcal{X}$.

Obliczenie wartości wyrażenia (5.2) wymaga w metodzie MC wygenerowania losowej próby $X_1, X_2, \ldots, X_n$ zmiennych *iid* z rozkładu o gęstości $f(x)$ przy pomocy odpowiednich komputerowych algorytmów pseudolosowych. Naturalna średnia postaci

$$\hat{h}_f(X) = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \tag{5.4}$$

przybliża wtedy, zgodnie z mocnym prawem wielkich liczb, wartość oczekiwaną $\mathbb{E}_f h(X)$:

Twierdzenie 5.1. *Niech $Y_1, Y_2, \ldots, Y_n$ będą zmiennymi losowymi* iid *o wartości oczekiwanej μ i skończonej wariancji σ^2. Wtedy*

$$\frac{\sum_{i=1}^{n} Y_i}{n} \xrightarrow[n\to\infty]{p.n.} \mu \ . \tag{5.5}$$

Wystarczy zatem przyjąć $h(X_i) = Y_i$ oraz $\mathbb{E}_f h(X) = \mu$, aby otrzymać szukaną zbieżność (5.4) do wartości oczekiwanej.

Metodę powyższą określa się czasami w literaturze jako surowe Monte Carlo (lub proste Monte Carlo, w j. ang. *crude Monte Carlo*) i można ją uznać za protoplastę pozostałych algorytmów (patrz [26]).

Zauważmy, że w tej wersji twierdzenia 5.1 istotną rolę gra skończoność wariancji pojedynczej zmiennej Y_i. W przypadku symulacji brak spełnienia założenia o skończoności σ^2 może prowadzić do ekstremalnej niestabilności estymatora (5.4), tzn. do nieprzewidywalnych zmian jego wartości pomiędzy kolejnymi trajektoriami symulacji.

Obecnie istnieje wiele różnorakich rozwinięć metody *crude* Monte Carlo, których głównym celem jest minimalizacja wariancji estymatora $\hat{h}_f(X)$. Do estymatora (5.4) można bowiem zastosować centralne twierdzenie graniczne.

Twierdzenie 5.2. *Niech $Y_1, \ldots, Y_n$ będą zmiennymi losowymi* iid *o wartości oczekiwanej μ i skończonej wariancji σ^2. Wtedy*

$$\frac{\sum_{i=1}^{n}(Y_i - \mu)}{\sigma\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{D}} N(0;1) \ . \tag{5.6}$$

W przypadku estymatora (5.4), z (5.6) mamy

$$\sqrt{n}\left(\hat{h}_f(X) - \mathbb{E}_f h(X)\right) \xrightarrow[n\to\infty]{\mathcal{D}} N\left(0; \mathrm{Var}(\hat{h}_f(X))\right) \ , \tag{5.7}$$

gdzie wariancję $\mathrm{Var}(\hat{h}_f(X))$ estymatora (5.4) możemy przybliżać naturalnym wzorem

$$\hat{\mathrm{Var}}(\hat{h}_f(X)) = \frac{1}{n}\sum_{i=1}^{n}(h(X_i) - \hat{h}_f(X))^2 \ . \tag{5.8}$$

Ze wzoru (5.7) wynika, iż redukcja wariancji estymatora $\hat{h}_f$ zwiększa dokładność uzyskiwanych wyników, pozwalając na zmniejszenie niezbędnej liczby losowań n. Prowadzi to oczywiście do skrócenie nieraz bardzo czasochłonnych symulacji.

Jak bardzo ten problem jest istotny, zobaczyć możemy na klasycznym przykładzie zaczerpniętym z [30].

Przykład 5.3. *Załóżmy, że interesuje nas obliczenie całki postaci*

$$I = \int_2^\infty \frac{1}{\pi(1+x^2)} dx \ . \tag{5.9}$$

Po pierwsze, zauważyć możemy, że całka ta jest podobna do gęstości standardowego rozkładu Cauchy'ego o postaci

$$f(t) = \frac{1}{\pi(1+t^2)} \ . \tag{5.10}$$

Zatem, jeśli dysponujemy generatorem z rozkładu Cauchy'ego, odpowiedni estymator metody MC dany może być formułą

$$\hat{I}_1 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1} \left(X_i > 2 \right) \tag{5.11}$$

dla $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$ *Cauchy. Odpowiedni generator dla rozkładu Cauchy'ego otrzymać można np. metodą ROU. Dla takiego estymatora mamy z rozkładu dwumianowego*

$$\operatorname{Var} \hat{I}_1 = \frac{I(1-I)}{n} \approx \frac{0,127}{n} \ . \tag{5.12}$$

Biorąc pod uwagę, że gęstość rozkładu Cauchy'ego jest symetryczna względem zera, możemy skonstruować inny estymator dla całki I *o postaci*

$$\hat{I}_2 = \frac{1}{2n} \sum_{i=1}^{n} \mathbb{1} \left(|X_i| > 2 \right) \ . \tag{5.13}$$

Zmniejsza to błąd estymacji, ponieważ

$$\operatorname{Var} \hat{I}_2 = \frac{I(1-2I)}{2n} \approx \frac{0,052}{n} \ . \tag{5.14}$$

Wynik ten można jeszcze poprawić. W tym celu należy zauważyć, że w tak zapisanej całce I *interesujemy się głównie „mało prawdopodobnymi" wartościami z rozkładu Cauchy'ego, tzn. wartościami większymi niż 2. Tymczasem główna masa prawdopodobieństwa w rozkładzie Cauchy'ego lokuje się wokół zera, a nie w ogonie. Dlatego też, jeśli dysponujemy odpowiednią informacją, całkę* (5.9) *możemy zapisać w innej formie*

$$I = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)} dx = \frac{1}{2} - \int_0^2 \frac{2}{\pi(1+x^2)} \frac{1}{2} dx \ . \tag{5.15}$$

Dla całki (5.15) *odpowiedni estymator ma postać*

$$\hat{I}_3 = \frac{1}{2} - \frac{1}{n} \sum_{i=1}^{n} \frac{2}{\pi(1+U_i^2)} \ , \tag{5.16}$$

gdzie $U_1, U_2, \ldots, U_n \overset{iid}{\sim} U[0;2]$. *Korzystając z całkowania przez części, otrzymujemy dla takiego estymatora*

$$\operatorname{Var} \hat{I}_3 \approx \frac{0,0092}{n} \ . \tag{5.17}$$

W przypadku całki (5.9) *zastosować możemy również podstawienie* $y = x^{-1}$. *Daje to alternatywny wzór*

$$I = \int_0^{\frac{1}{2}} \frac{1}{\pi(1+\frac{1}{y^2})y^2} dy = \int_0^{\frac{1}{2}} \frac{1}{2\pi(1+y^2)} 2dy \ , \tag{5.18}$$

który prowadzi do estymatora postaci

$$\hat{I}_4 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2\pi(1+U_i^2)} \ , \tag{5.19}$$

gdzie $U_1, U_2, \ldots, U_n \overset{iid}{\sim} U[0;\frac{1}{2}]$. *W takim przypadku, poprzez całkowanie przez części, otrzymujemy*

$$\operatorname{Var} \hat{I}_4 \approx \frac{0,00095}{n} \ . \tag{5.20}$$

Porównując (5.12) *z* (5.20)*, widzimy, że udało nam się zmniejszyć wariancję estymatora około 1000 razy, co daje* $\sqrt{1000} \approx 33$ *razy mniej symulacji wymaganych do otrzymania tej samej dokładności. Ceną jest zwiększanie się ilości wiedzy i nakładu czasu poświęconego na konstrukcję estymatora.*

Przykładami metod pozwalających uzyskać redukcję wariancji jest próbkowanie ważone (w j. ang. *importance sampling*), metoda zmiennych antytetycznych (w j. ang. *antithetic variables*), czy zmiennych kontrolnych (w j. ang. *control variates*). Pokrótce omówimy teraz te zagadnienia.

Metoda próbkowania ważonego polega na zaobserwowaniu, że

$$\mathbb{E}_f h(X) = \int_{\mathcal{X}} h(x) \frac{f(x)}{g(x)} g(x) dx \ , \tag{5.21}$$

czyli szukany estymator $\hat{h}_f(X)$ możemy obliczyć z alternatywnego wzoru

$$\hat{h}_f(X) = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} h(X_i) \ , \tag{5.22}$$

gdzie $X_1, \ldots, X_n$ są zmiennymi *iid* generowanymi z rozkładu o gęstości $g(x)$. Estymator (5.22) zbiega do $\mathbb{E}_f h(X)$ przy omawianych wcześniej założeniach oraz dodatkowym, iż nośnik gęstości $g(x)$ zawiera nośnik $f(x)$.

Przy odpowiednim wyborze gęstości $g(x)$, możemy zmniejszyć wariancję estymatora $\operatorname{Var}(\hat{h}_f(X))$. O optymalnym wyborze tej gęstości mówi następujące twierdzenie.

Twierdzenie 5.4. *Wariancja estymatora* (5.22) *jest minimalna dla gęstości postaci*

$$g^*(x) = \frac{|h(x)|f(x)}{\int_{\mathcal{X}} |h(z)|f(z)dz} \ . \tag{5.23}$$

Dowód powyższego twierdzenia znaleźć można np. w [32]. Niestety, nie jest możliwe bezpośrednie zastosowanie wzoru (5.23), gdyż postać $g^*(x)$ zależy w nim od *szukanej* wartości całki (5.2). Powyższe twierdzenie jest jednak przydatne w praktyce, gdyż przy posiadaniu nawet tylko częściowych informacji dotyczących postaci prawej strony (5.23), umożliwia ono dobranie odpowiedniej funkcji gęstości $g^*(x)$, gwarantującej szybszą zbieżność algorytmu.

W przypadku metody zmiennych antytetycznych zamiast standardowego estymatora (5.4) wykorzystujemy estymator postaci

$$\hat{h}_f(X) = \frac{1}{2n} \sum_{1=0}^{n} \left(h(X_i) + h(Y_i) \right) \ , \tag{5.24}$$

gdzie *pary* zmiennych $(X_1, Y_1), \ldots (X_n, Y_n)$ są parami zmiennych *iid.* Zarówno $X_1, \ldots, X_n$, jak i $Y_1, \ldots, Y_n$ pochodzą przy tym z gęstości $f(x)$. Co jednak istotne, wartości zmiennych w każdej z par (X_i, Y_i) są generowane algorytmem pseudolosowym w specjalny sposób – tak, aby $h(X_i)$ i $h(Y_i)$ były wzajemnie ujemnie skorelowane. Zmienne Y_i nazywane są wtedy zmiennymi antytecznymi (względem zmiennych X_i). W oczywisty sposób, dzięki ujemnej korelacji, wariancja estymatora (5.24) jest niższa od wariancji estymatora standardowego (5.4), nawet przy podwojonej długości tego ostatniego.

Metoda zmiennych kontrolnych zakłada znajomość dwóch różnych estymatorów. Pierwszy z nich to $\hat{h}_f^{(1)}(X)$, estymujący szukaną wartość oczekiwaną $\mathbb{E}_f h(X)$. Drugi zaś – $\hat{h_0}_f^{(2)}(X)$ – estymuje wartość oczekiwaną pewnej innej funkcji $h_0(x)$. Ponieważ w oczywisty sposób

$$\mathbb{E}_f h(X) = (\mathbb{E}_f h(X) + c\mathbb{E}_f h_0(X)) - c\mathbb{E}_f h_0(X) \ , \tag{5.25}$$

więc dysponując $\hat{h}_f^{(1)}(X)$ i $\hat{h_0}_f^{(2)}(X)$, możemy stworzyć nowy estymator, „łączący" informacje dostarczane przez oba te estymatory. W szczególności, jeśli

$$\hat{h}_f^{(1)}(X) = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \ , \ \hat{h_0}_f^{(2)}(X) = \frac{1}{n} \sum_{i=1}^{n} h_0(X_i) \ , \tag{5.26}$$

to estymator zbudowany na ich podstawie jest postaci

$$\hat{h}_f(X) = \frac{1}{n} \sum_{i=1}^{n} \left(h(X_i) + c h_0(X_i) \right) - c\mathbb{E}_f h_0(X) \ . \tag{5.27}$$

Jak widzimy, w istotny sposób w (5.27) zakładamy *znajomość* wartości wielkości $\mathbb{E}_f h_0(X)$. Wariancja wspólnego estymatora (5.27) jest równa

$$\begin{aligned} \mathrm{Var}(\hat{h}_f(X)) = \mathrm{Var}\left(\hat{h}_f^{(1)}(X) \right) + c^2 \mathrm{Var}\left(\hat{h_0}_f^{(2)}(X) \right) + \\ + 2c \, \mathrm{Cov}\left(\hat{h}_f^{(1)}(X), \hat{h_0}_f^{(2)}(X) \right) \ . \end{aligned} \tag{5.28}$$

Jeśli przyjmiemy optymalną wielkość stałej c wynoszącą

$$c^* = - \frac{\mathrm{Cov}\left(\hat{h}_f^{(1)}(X), \hat{h_0}_f^{(2)}(X) \right)}{\mathrm{Var}\left(\hat{h_0}_f^{(2)}(X) \right)} \ , \tag{5.29}$$

to zminimalizuje ona całkowitą wariancję (5.28). Jak widać, metoda zmiennych kontrolnych polega na wykorzystaniu dodatkowej funkcji $h_0(x)$ jako źródła zwiększenia ilości informacji na temat estymatora (5.27), co może skutkować zmniejszeniem wariancji estymatora $\hat{h}_f(X)$.

Szersze informacje na temat wyżej wzmiankowanych metod znaleźć można np. w [30].

5.2 Zagadnienie optymalizacji metodą MC

Drugą klasą problemów, które można rozwiązać poprzez zastosowanie metod MC, są zagadnienia optymalizacji, ze szczególnym uwzględnieniem kwestii poszukiwania ekstremów globalnych danej funkcji. Innymi słowy, zainteresowani jesteśmy rozwiązaniem problemu

$$\max_{x\in\mathcal{X}} h(x) \tag{5.30}$$

dla pewnej funkcji $h(x)$ o dziedzinie $\mathcal{X} \subset \mathbb{R}^p$.

Najprostszym rozwiązaniem problemu (5.30) przy założeniu zwartości przestrzeni stanów $\mathcal{X}$ jest wygenerowanie próby $X_1, X_2, \ldots, X_n$ z rozkładu jednostajnego na $\mathcal{X}$. Wtedy trywialny estymator

$$\hat{h}_{\max} = \max_{i=1,\ldots,n} \{h(X_i)\} \tag{5.31}$$

jest naturalnym estymatorem rozwiązania dla zadania optymalizacji (5.30).

Oczywiście także i dla problemu (5.30), podobnie jak dla problemu całkowania funkcji, możliwe jest zastosowanie szeroko znanych, deterministycznych metod numerycznych (chociażby np. metody Newtona, metody siecznych czy ich uogólnień – patrz np. [12]). Jednak jeśli rozpatrywana funkcja $h(x)$ posiada kilka maksimów lokalnych, zastosowanie metody deterministycznej może zaowocować „uwięźnięciem" algorytmu, tzn. znalezieniem tylko jednego z ekstremów lokalnych, a nie prawdziwego maksimum globalnego. Ponadto w przypadku deterministycznych metod numerycznych częstokroć konieczne są dodatkowe założenia co do szczególnych własności funkcji $h(x)$, jak np. jej różniczkowalność, lub co do odpowiednio regularnej postaci jej dziedziny $\mathcal{X}$.

5.2.1 Symulowane wyżarzanie

Remedium na wspomnianie wyżej problemy może być zastosowanie metody MC zwanej symulowanym wyżarzaniem (w j. ang. *simulated annealing*) z modyfikacją zaproponowaną przez Metropolisa (patrz np. [15, 25, 30]).

Metoda ta wymaga wprowadzenia dodatkowej zmiennej skalującej $T > 0$, tradycyjnie nazywanej *temperaturą*, ze względu na historyczne zastosowania symulowanego wyżarzania w fizyce. Startując z pewnej wartości x_0, metoda polega na generacji kolejnych wartości X_n z wykorzystaniem następującego algorytmu:

Algorytm 5.5.

1. Wylosuj *proponowaną wartość* Y_n z pewnego rozkładu prawdopodobieństwa na otoczeniu wartości poprzedniego kroku $X_{n-1} = x_{n-1}$, np. z rozkładu jednostajnego na otoczeniu tego punktu. W ogólności losowanie nastąpić może z dowolnego rozkładu o gęstości postaci $g(|x_{n-1} - Y_n|)$).

2. Wybierz kolejny punkt X_n według następującego przepisu:

$$X_n = \begin{cases} Y_n & \text{z prawd. } p = \min\{\exp(\Delta h/T), 1\} \\ x_{n-1} & \text{z prawd. } 1-p \end{cases} , \tag{5.32}$$

gdzie $\Delta h = h(Y_n) - h(x_{n-1})$.

Jak widzimy z (5.32), jeśli funkcja $h(x)$ ma większą wartość w nowo wylosowanym punkcie Y_n, to zostanie on na pewno zaakceptowany i stanie się zmienną X_n. Jednocześnie, nawet gdy wartość funkcji jest niższa w nowym punkcie Y_n, to punkt ten może zostać zaakceptowany na X_n z niezerowym prawdopodobieństwem $p = \exp(\Delta h/T)$. Umożliwia to ucieczkę z ewentualnego maksimum lokalnego i daje szansę na znalezienie maksimum globalnego.

Powyższy algorytm jest zazwyczaj implementowany z dodatkowym warunkiem monotonicznej zbieżności $T \to 0$. Zmniejszanie się parametru temperatury ma celu stopniowe „zamrażanie" kroków algorytmu, który najpierw szybko przeszukuje całą dziedzinę, aby stopniowo zatrzymać się w ekstemum. Jak z tego wynika, na odpowiednie zachowanie się procedury, a więc i na znalezienie z dużym prawdopodobieństwem globalnego ekstremum, olbrzymi wpływ ma sposób zbiegania parametru temperatury T do zera.

Należy również podkreślić, że przedstawiony powyżej algorytm nie jest już typową procedurą Monte Carlo, należy raczej do dziedziny metod MCMC (Markov Chain Monte Carlo).

5.2.2 Metoda EM

Nazwa algorytmu EM pochodzi od dwóch kroków wykorzystywanych w tej metodzie – kroku obliczania wartości oczekiwanej (krok E – *Expectation*) i kroku szukania wartości maksymalnej (krok M – *Maximization*). Zanim jednak przystąpimy do opisania algorytmu EM, pokrótce przedstawimy modele *brakującej zmiennej*, w których metoda ta może być wykorzystana.

W przypadku modeli brakującej zmiennej (*missing data models*), funkcja, którą chcemy optymalizować, może zostać przedstawiona w postaci

$$h(x) = \mathbb{E}_Z\left(H(x, Z)\right) , \tag{5.33}$$

czyli jako wartość oczekiwana obliczana względem dodatkowej zmiennej Z. Zazwyczaj zmienna Z wynika bezpośrednio z rozważanego modelu (np. w przypadku danych cenzorowanych, gdzie nasza informacja ograniczona jest przez istnienie dodatkowej zmiennej cenzurującej) lub może być dodana sztucznie, w celu ułatwienia rozwiązania problemu optymalizacji.

Model (5.33) rozważany może być na gruncie statystyki, ze szczególnym uwzględnieniem pojęcia wiarygodności i estymacji metodą największej wiarygodności (patrz rozdział 1.2.12). W takim przypadku funkcją wiarygodności pełnego modelu (*complete-model likelihood*) nazywamy

$$L^c\left(\theta|x, z\right) = f\left(x, z|\theta\right) = f_\theta\left(x, z\right) , \tag{5.34}$$

gdzie $f\left(x, z|\theta\right)$ jest łączną gęstością zmiennych X i Z przy ustalonym parametrze modelu statystycznego θ.

W takim modelu zakładamy, że obserwujemy próbę $X_1, X_2, \ldots, X_n$ zmiennych *iid* z rozkładu o gęstości $g(x|\theta) = g_\theta(x)$. Interesuje nas zadanie estymacji metodą największej wiarygodności, czyli znalezienie

$$\hat{\theta} = \sup_\theta L(\theta) = \sup_\theta L(\theta|x_1, x_2, \ldots, x_n) = \sup_\theta L(\theta|\mathbf{x}) \ , \tag{5.35}$$

gdzie $\mathbf{x}$ będzie w skrócie oznaczać wektor znanych nam wartości obserwacji $x_1, x_2, \ldots, x_n$. Obserwacje dodatkowych zmiennych $Z_1, Z_2, \ldots, Z_n$ będziemy analogicznie oznaczać przez $\mathbf{z} = z_1, z_2, \ldots, z_n$. Łączna gęstość $X_1, \ldots, X_n$ i $Z_1, \ldots, Z_n$ przy ustalonym parametrze θ jest postaci $f(\mathbf{x}, \mathbf{z}|\theta) = f_\theta(\mathbf{x}, \mathbf{z})$.

Z definicji gęstości warunkowej mamy

$$k(z|\theta, x) = \frac{f(x, z|\theta)}{g(x|\theta)} \ , \tag{5.36}$$

gdzie $k(z|\theta, x)$ jest gęstością warunkową brakującej zmiennej Z pod warunkiem parametru θ i dla znanej zmiennej X. Z (5.36) otrzymujemy relację łączącą wiarygodność pełnego modelu i wiarygodność $L(\theta|\mathbf{x})$ dla zmiennych obserwowanych

$$\log L(\theta|\mathbf{x}) = \mathbb{E}_{k,\theta_0} \left(\log L^c \left(\theta|\mathbf{x}, \mathbf{z} \right) \right) - \mathbb{E}_{k,\theta_0} \left(k(\mathbf{z}|\theta, \mathbf{x}) \right) \ , \tag{5.37}$$

gdzie wartość oczekiwana liczona jest względem ustalonej wartości parametru θ_0 i gęstości $k(\mathbf{z}|\theta_0, \mathbf{x})$. Co istotne, w celu znalezienia maksimum $\log L(\theta|\mathbf{x})$ wystarczy skupić się na pierwszym elemencie po prawej stronie równania (5.37). Jeśli wprowadzimy oznaczenie

$$Q\left(\theta|\theta_0, \mathbf{x}\right) = \mathbb{E}_{k,\theta_0} \left(\log L^c \left(\theta|\mathbf{x}, \mathbf{z} \right) \right) \ , \tag{5.38}$$

otrzymujemy następujący iteracyjny algorytm, startujący z ustalonej wartości θ_0:

Algorytm 5.6.

1. (Krok E) Dla ustalonej wartości θ_n, oblicz

$$Q\left(\theta|\theta_n, \mathbf{x}\right) = \mathbb{E}_{k,\theta_n} \left(\log L^c \left(\theta|\mathbf{x}, \mathbf{z} \right) \right) \tag{5.39}$$

2. (Krok M) Znajdź

$$\theta_{n+1} = \arg\max_\theta Q\left(\theta|\theta_n, \mathbf{x}\right) \ . \tag{5.40}$$

Zbieżność takiego algorytmu jest zapewniana przez następujące twierdzenia.

Twierdzenie 5.7. *Ciąg $\theta_0, \theta_1, \ldots$ wygenerowany algorytmem EM spełnia warunek*

$$L(\theta_{n+1}|\mathbf{x}) \geqslant L(\theta_n|\mathbf{x}) \ , \tag{5.41}$$

przy czym równość zachodzi wtedy i tylko wtedy, gdy

$$Q\left(\theta_{n+1}|\theta_n, \mathbf{x}\right) = Q\left(\theta_n|\theta_n, \mathbf{x}\right) \ . \tag{5.42}$$

Twierdzenie 5.8. *Jeśli funkcja $Q\left(\theta|\theta_n, \mathbf{x}\right)$ jest ciągła zarówno dla θ, jak i θ_n, wtedy każda granica ciągu $\theta_0, \theta_1, \ldots$ jest punktem stacjonarnym $L(\theta|\mathbf{x})$, a ciąg $L(\theta_0|\mathbf{x}), L(\theta_1|\mathbf{x})$ zbiega monotonicznie do $L(\hat{\theta}|\mathbf{x})$ dla pewnego $\hat{\theta}$.*

Jak widzimy, twierdzenie 5.7 oznacza, że wartość funkcji wiarygodności zwiększa się wraz z każdym krokiem algorytmu EM. Z kolei z twierdzenia 5.8 wynika, że istnieje zbieżność, ale niekoniecznie do globalnego maksimum, czyli naszego poszukiwanego estymatora w problemie (5.35). Z tego względu stosuje się dodatkowe techniki zapewniające zbieżność do maksimum globalnego, jak np. rozpoczynanie metody EM z różnych punktów startowych.

Metoda MC może być zastosowana do algorytmu EM, dając tzw. metodę MCEM (*Monte Carlo EM*). Stosuje się ją w kroku E, ze względu na problemy powstające przy obliczaniu wartości oczekiwanej (5.39). Remedium może być generowanie próby $Z_1, Z_2, \ldots, Z_m$ zmiennych z rozkładu o gęstości warunkowej $k(\mathbf{z}|\theta_n, \mathbf{x})$. Wtedy estymator postaci

$$\hat{Q}\left(\theta|\theta_n, \mathbf{x}\right) = \frac{1}{m}\sum_{i=1}^{m} \log L^c\left(\theta|\mathbf{x}, Z_1, Z_2, \ldots, Z_m\right) \tag{5.43}$$

przybliża odpowiednią wartość w kroku E.

5.3 Zastosowania i ograniczenia metod MC

Niektóre z zastosowań metod MC zostały wymienione na początku tego rozdziału. Tutaj pokrótce przyjrzymy się tylko wykorzystaniu metody *crude Monte Carlo* (patrz wzór (5.4)) w wycenie pochodnych instrumentów finansowych.

W klasycznym modelu Blacka – Scholesa jednym z założeń jest, iż trajektoria podstawowego instrumentu finansowego opisywana jest *geometrycznym ruchem Browna o stałym dryfie μ i zmienności σ* (patrz np. [18])

$$S_t = S_0 \exp(\mu t + \sigma \mathcal{W}_t) \ , \tag{5.44}$$

gdzie S_t jest ceną podstawowego instrumentu finansowego w chwili $t \in [0; T]$, $\mathcal{W}_t$ jest standardowym (arytmetycznym) ruchem Browna. Po zastosowaniu równoważnej miary martyngałowej (patrz np. [36]) otrzymujemy przekształcenie (5.44) o postaci

$$S_t = S_0 \exp\left(\left(r - \frac{1}{2}\sigma^2\right)t + \sigma \mathcal{W}_t\right) \ , \tag{5.45}$$

gdzie r jest stałą stopą procentową w okresie $[0; T]$. Jeśli dla przykładu założymy, że podstawowym instrumentem finansowym jest *obligacja o chwili zapadalności T*, a związanym z nią pochodnym instrumentem finansowym tzw. *europejska opcja call* (patrz np. [18]) o funkcji wypłaty danej wzorem

$$\max\{0; S_T - K\} \ , \tag{5.46}$$

gdzie K jest tzw. *ceną wykonania opcji*, to cena tego typu opcji opisana jest wzorem

$$C^* = \mathbb{E}\left(e^{-rT} \max\{0; S_T - K\}\right) \ . \tag{5.47}$$

Jak widzimy, w (5.47) występuje wartość oczekiwana *zdyskontowanych przyszłych strumieni płatności* (patrz np. [18, 36]). Jeśli przez $\mathcal{S}_T^{(1)}, \mathcal{S}_T^{(2)}, \ldots, \mathcal{S}_T^{(n)}$ oznaczymy n wygenerowanych zgodnie ze wzorem (5.45) wartości cen instrumentu podstawowego w chwili T, to estymatorem ceny opcji C^* jest

$$\hat{C}^* = \frac{1}{n}\sum_{i=1}^{n} e^{-rT} \max\{0; \mathcal{S}_T^{(i)} - K\} \ , \tag{5.48}$$

czyli klasyczny estymator metody *crude Monte Carlo*. Dalsze szczegóły odnośnie wykorzystania metod Monte Carlo w finansach znaleźć można np. w [16].

Innym ważnym przykładem są zastosowania statystyczne metod MC, np. w testach statystycznych. Przy rozpatrywaniu testu proporcji (zwanego też testem frakcji) w przykładzie 1.29, w celu ułatwienia obliczeń przeszliśmy od rozkładu dwumianowego do prostszego, bo stablicowanego rozkładu normalnego. Oczywiście możemy też postąpić w inny sposób – spróbować zdefiniować obszar krytyczny w teście na podstawie *symulacji*. W przykładzie z monetą, możemy wiele razy wygenerować *scenariusze* $W_1, W_2, \ldots, W_m$ wyników rzutu 1000 razy monetą. Każdy z takich scenariuszy W_i polegać będzie na 1000 razy powtarzanym prostym schemacie – wylosuj „reszkę" lub „orła" z równymi prawdopodobieństwami 0,5. Następnie dla każdego scenariusza zliczamy liczbę wystąpień orłów X_i. Do obszaru krytycznego wybieramy wszystkie scenariusze $|X_i - 500| > c$, tzn. takie, dla których liczba orłów jest albo „zbyt mała" albo „zbyt duża". Liczba tych scenariuszy powinna wynosić $\alpha \cdot m$, gdzie α jest poziomem istotności naszego testu. Uzyskana w ten sposób wartość c opisuje nam odpowiedni obszar krytyczny K.

Bardzo podobna metodologia może zostać zastosowana do zagadnień związanych z obrotem i wyceną specyficznych instrumentów finansowych zwanych obligacjami katastroficznymi.

Jak wcześniej wspomniano, w przypadku zagadnień całkowania i optymalizacji można, oprócz metod Monte Carlo, zastosować także deterministyczne metody numeryczne. Metody symulacyjne posiadają kilka zalet w stosunku do metod deterministycznych, np. algorytm symulowanego wyżarzania posiada niezerowe prawdopodobieństwo opuszczenia swego lokalnego maksimum (patrz rozdział 5.2). Znacznie istotniejszą jednak zaletą metod Monte Carlo jest kwestia wspomnianego wcześniej tzw. *curse of dimensionality* (przekleństwa wymiaru). Dla deterministycznych metod błąd całkowania np. przez sumy Riemanna, jest rzędu $\mathcal{O}(m^{-4/d})$, gdzie m jest w tym przypadku liczbą punktów próbkowania, a d jest liczbą zmiennych całkowanej funkcji, czyli wymiarów przestrzeni. Tymczasem dla metody Monte Carlo analogiczny błąd jest rzędu $\mathcal{O}(m^{-1})$. Oznacza to, że dla liczby wymiarów większej niż 4, metoda Monte Carlo posiada błąd mniejszy niż deterministyczne metody numeryczne (patrz np. [30, 38]).

Jednym z poważniejszych ograniczeń metody Monte Carlo jest konieczność posiadania narzędzia, które pozwala na generowanie próbek *iid* z gęstości $f(x)$ dla estymatora (5.4) lub $g(x)$ dla (5.22).

Rozdział 6

Metody Markov Chain Monte Carlo

W największym skrócie rzecz ujmując, metody Markov Chain Monte Carlo (w skrócie MCMC) są rezultatem połączenia symulacji metodami Monte Carlo z teorią łańcuchów Markowa. Celem tego połączenia jest wyeliminowanie ograniczenia metod Monte Carlo wspomnianego w rozdziale 5.3 – konieczności generowania próbek *iid bezpośrednio* z pewnej ustalonej funkcji gęstości. Gęstość ta może mieć bowiem zbyt skomplikowaną postać, aby wykorzystanie jej było efektywne numerycznie lub nawet w ogóle możliwe. Może także być znana jedynie z dokładnością do pewnej stałej normującej, a stała ta, szczególnie w zagadnieniach fizycznych, jest zbyt skomplikowana do dostatecznie szybkiego obliczenia (patrz np. [5, 15, 30]).

W metodach MCMC zamiast próbek *iid* pochodzących z rozkładu prawdopodobieństwa o gęstości $f(x)$ wykorzystujemy losowy ciąg $(X_i)_{i=1}$ będący ŁM o rozkładzie stacjonarnym o gęstości $f(x)$. Dzięki temu unikamy w ogóle konieczności stworzenia odpowiedniego algorytmu dla generowania próbek z potencjalnie skomplikowanej funkcji gęstości $f(x)$. Zamiast tego, dzięki zastosowaniu twierdzeń ergodycznych dla ŁM (patrz rozdział 1.6.3) możemy skorzystać z pewnych ogólnych algorytmów bazujących jedynie na odpowiednich prawdopodobieństwach przejścia pomiędzy poszczególnymi stanami ŁM.

Najprostszą ilustracją tego zagadnienia jest zastosowanie metody MCMC w zadaniu (5.2) całkowania iloczynu funkcji $h(x)f(x)$, które polega na wygenerowaniu ŁM $(X_i)_{i=1}$ o rozkładzie stacjonarnym zadanym gęstością $f(x)$. Wtedy bazujący na tym ciągu klasyczny estymator (5.4) jest także estymatorem wartości szukanej całki zgodnie z wnioskiem 1.68 ze słabego twierdzenia ergodycznego dla ŁM.

Do stworzenia odpowiedniej próby $X_1, X_2, \ldots, X_n, \ldots$, która będzie jednocześnie łańcuchem Markowa o z góry zadanym rozkładzie stacjonarnym, służą dwa bardzo ogólne algorytmy, zwane algorytmem Metropolisa – Hastingsa i próbnikiem Gibbsa (w j. ang. *Gibbs sampler*). Przyjrzymy im się teraz dokładniej.

6.1 Algorytm Metropolisa – Hastingsa

W celu wygenerowania ciągu $X_1, X_2, \ldots, X_n, \ldots$ w algorytmie Metropolisa – Hastingsa (w skrócie określanym jako algorytm MH) wykorzystywana jest dodatkowa, ustalana przez eksperymentatora, gęstość warunkowa $g(y|x)$, zwana *gęstością proponującą* lub *instrumentalną* (w j. ang. *proposal* lub *instrumental denstity*). Gęstość ta może być wybrana dość dowolnie, jednak w celu zapewnienia sprawnego i poprawnego działania algorytmu HM, powinna spełniać poniższe warunki:

1. Gęstość $g(y|x)$ musi mieć postać łatwą do generowania z niej próbek względem zmiennej y dla każdego ustalonego x.
2. Musi istnieć możliwość obliczenia wielkości $f(y)/g(y|x)$, ewentualnie z dokładnością do stałej niezależnej od zmiennej x, lub gęstość instrumentalna powinna być symetryczna, tzn. spełniać warunek $g(x|y) = g(y|x)$.
3. Rodzina nośników funkcji $g(.|x)$ powinna zawierać cały nośnik gęstości $f(.)$.
4. Powstały ŁM powinien być nieprzywiedlny, nieokresowy i powracający w sensie Harrisa. Wystarczające ku temu warunki sformułujemy za chwilę.

Startując z dowolnej, początkowej wartości x_0, algorytm MH generuje kolejne wyrazy łańcucha Markowa poprzez wykonanie serii następujących kroków

Algorytm 6.1.

1. Wylosowanie zmiennej Y_{i-1} z gęstości $g(.|x_{i-1})$
2. Nowy punkt X_i losowany jest według przepisu:

$$X_i = \begin{cases} Y_{i-1} & \text{z prawd. } p(x_{i-1}, Y_{i-1}) \\ x_{i-1} & \text{z prawd. } 1 - p(x_{i-1}, Y_{i-1}) \end{cases} , \tag{6.1}$$

gdzie

$$p(x, y) = \min \left\{ \frac{f(y)}{g(y|x)} \frac{g(x|y)}{f(x)}, 1 \right\} . \tag{6.2}$$

Prawdopodobieństwo $p(x, y)$ nazywane jest *prawdopodobieństwem akceptacji* (w j. ang. *acceptance probability*).

Zauważmy duże podobieństwa pomiędzy przedstawionym powyżej algorytmem a symulowanym wyżarzaniem (patrz rozdział 5.2). W obu algorytmach najpierw losowany jest punkt – kandydat z otoczenia poprzedniej wartości zgodnie z dodatkową, niezależną gęstością, który zostaje wartością nowego kroku z pewnym ustalonym prawdopodobieństwem p. Prawdopodobieństwo akceptacji p jest przy tym zależne od interesującej nas funkcji, odpowiednio – optymalizowanej funkcji $h(x)$ dla symulowanego wyżarzania, lub gęstości $f(x)$ w algorytmie MH. Podobieństwo to jest szczególnie uderzające, jeśli w algorytmie MH zastosujemy symetryczną gęstość instrumentalną $g(y|x)$. Wtedy prawdopodobieństwo akceptacji przyjmuje postać

$$p(x, y) = \min \left\{ \frac{f(y)}{f(x)}, 1 \right\} . \tag{6.3}$$

Należy podkreślić, że w algorytmie MH nie następuje generowanie zmiennych losowych z gęstości $f(x)$, a jedynie z gęstości instrumentalnej $g(y|x)$, której postać jest właściwie niezależna od $f(x)$. Jedyna informacja, która jest niezbędna do wyliczenia wartości prawdopodobieństwa akceptacji to stosunek wartości $\frac{f(y)}{f(x)}$. W szczególności oznacza to, że nie jest niezbędna znajomość czynnika normującego dla gęstości $f(x)$.

Dla algorytmu MH zachodzi istotne twierdzenie.

Twierdzenie 6.2. *Gęstość $f(x)$ jest gęstością stacjonarną dla ŁM wygenerowanego zgodnie z algorytmem MH dla dowolnej gęstości instrumentalnej $g(y|x)$, której nośnik zawiera przestrzeń $\mathcal{X}$.*

Dowód powyższego twierdzenia znaleźć można np. w [30].

Zajmiemy się teraz problemem zbieżności ciągu wygenerowanego algorytmem MH do gęstości stacjonarnej $f(x)$ jako zbieżności granicznej, co umożliwi zastosowanie twierdzeń ergodycznych z rozdziału 1.6.3. Jak było to zaznaczone przy odpowiednich twierdzeniach, interesujący nas ŁM powinien być nieprzywiedlny, nieokresowy i powracający w sensie Harrisa. W przypadku algorytmu MH warunki te zachodzą, jeśli spełnione są łatwe do sprawdzenia w praktyce zależności.

Przytoczymy teraz przykłady odpowiednich twierdzeń tego dotyczących.

Twierdzenie 6.3. *Jeśli dla dowolnych $x, y \in \mathcal{X} \times \mathcal{X}$ zachodzi nierówność*

$$g(y|x) > 0 \ , \tag{6.4}$$

to ŁM generowany przy pomocy algorytmu MH jest nieprzywiedlny i powracający w sensie Harrisa. Jeśli dodatkowo spełniony jest warunek

$$\mathrm{P}\left(\frac{f(y)}{g(y|x)}\frac{g(y|x)}{f(x)} \leqslant 1\right) < 1 \ , \tag{6.5}$$

to ten ŁM jest nieokresowy.

Dowód powyższego twierdzenia znaleźć można np. w [30].

Warunek (6.4) jest równoważny wymaganiu, aby możliwe było przejście pomiędzy dowolnymi dwoma stanami x i y dokładnie w jednym kroku. Z kolei nierówność (6.5) jest równoważna warunkowi

$$P(X_i = X_{i-1}) > 0 \tag{6.6}$$

czyli, że szansa pozostania ŁM w tym samym punkcie jest niezerowa.

Zamiast warunków przedstawionych w twierdzeniu 6.3 możliwe jest zastosowanie innego twierdzenia, również gwarantującego odpowiednie własności ŁM wygenerowanego algorytmem MH.

Twierdzenie 6.4. *Niech gęstość $f(x)$ będzie ograniczona i dodatnia na każdym zbiorze zwartym zawartym w przestrzeni $\mathcal{X}$. Jeśli istnieją stałe $\epsilon, \delta > 0$, t.ż.*

$$g(y|x) > \epsilon \ dla \ |x - y| < \delta \ , \tag{6.7}$$

to ŁM wygenerowany za pomocą algorytmu MH jest nieprzywiedlny, powracający w sensie Harrisa i nieokresowy.

Dowód powyższego twierdzenia znaleźć można np. w [30].

Warunek (6.7) oznacza, że możemy zminoryzować od dołu funkcję gęstości pomocniczej na dowolnie małym otoczeniu każdego stanu x. Intuicyjnie odpowiada to żądaniu, aby zarówno pozostanie w tym samym stanie, jak i przejście do pobliskiego stanu odbywało się z pewnym niezerowym prawdopodobieństwem.

Istnieją rozmaite możliwości wyboru kandydatów na gęstość instrumentalną $g(y|x)$. Oczywiście, w danym szczególnym przypadku może mieć ona prawie dowolną postać. Istnieją jednak pewne ogólne metody wyboru postaci gęstości proponującej, które teraz pokrótce omówimy.

Jedną z najprostszych postaci gęstości $g(y|x)$ jest funkcja niezależna od zmiennej x, tzn. gęstość $g(y)$. W tym przypadku odpowiedni algorytm MH ma postać

1. Wylosuj zmienną Y_{i-1} z gęstości $g(.)$
2. Nowy punkt X_i losuj według przepisu:

$$X_i = \begin{cases} Y_{i-1} & \text{z prawd. } p(x_{i-1}, Y_{i-1}) \\ x_{i-1} & \text{z prawd. } 1 - p(x_{i-1}, Y_{i-1}) \end{cases} , \tag{6.8}$$

gdzie

$$p(x, y) = \min \left\{ \frac{f(y)}{g(y)} \frac{g(x)}{f(x)}, 1 \right\} . \tag{6.9}$$

Jak widzimy, w tej modyfikacji algorytmu MH, znanej jako *niezależny algorytm MH* (w j. ang. *independent Metropolis-Hastings algorithm*), generowanie kolejnych zmiennych Y_i odbywa się *niezależnie* od wartości zmiennej X_i.

Inną możliwością jest wykorzystanie gęstości instrumentalnej, która byłaby losowym zaburzeniem wartości zmiennej X_i, tzn. miała postać

$$Y_i = X_i + \varepsilon_i , \tag{6.10}$$

gdzie ε_i jest pewną zmienną losową niezależną od zmiennej X_i. W przypadku (6.10) odpowiedni ŁM utworzony przez wartości wygenerowane gęstością instrumentalną będzie *błądzeniem przypadkowym* po przestrzeni stanów $\mathcal{X}$. Jak łatwo zauważyć, dla (6.10) gęstość warunkową $g(y|x)$ możemy zapisać także jako funkcję zależną jedynie od odległości pomiędzy zmiennymi Y_i i X_i, tzn. w postaci $g(y - x)$.

Kolejną propozycją, wspomnianą już wcześniej, jest symetryczna postać gęstości instrumentalnej, tzn. spełniająca warunek $g(y|x) = g(x|y)$ dla dowolnych $x, y \in \mathcal{X}$. Jak pamiętamy, symetryczność gęstości proponującej umożliwia znaczne uproszczenie prawdopodobieństwa akceptacji do postaci (6.3), a więc de facto niezależnej od gęstości warunkowej $g(y|x)$.

Należy zauważyć, że wybór postaci gęstości instrumentalnej jest bardzo istotny w algorytmie MH. Z jednej strony powinna być ona możliwie jak najprostsza w celu ułatwienia konstrukcji algorytmu i przyśpieszenia jego działania, z drugiej strony powinna ona sprzyjać szybkiej zbieżności estymatora (5.4) do szukanej całki w problemie (5.2).

W przypadku wyboru gęstości instrumentalnej $g(y)$ niezależnej od aktualnej wartości zmiennej X_i, najlepsze rezultaty daje postać gęstości maksymalizującej prawdopodobieństwo akceptacji (6.9), co jednak zakłada podobieństwo funkcji $g(.)$ do $f(.)$. Z kolei w przypadku omówionego powyżej błądzenia losowego,

ważnym czynnikiem przy wyborze postaci gęstości instrumentalnej powinna być szybkość przeszukiwania przestrzeni stanów $\mathcal{X}$. Niestety, w obu przypadkach zalecenia te mają raczej charakter ogólnych wskazówek niż dokładnych wzorów na postać funcji $g(y|x)$ (patrz np. [30]).

Oprócz naturalnego estymatora średniej (5.4) dla algorytmu MH możliwe jest stworzenie innych, bardziej złożonych estymatorów średniej $\mathbb{E}_f h(X)$. Przykładem może być tutaj tzw. estymator *warunkowy* lub inaczej *Rao-Blackwellizowany* (w j. ang. *Rao-Blackwellized*). Co istotne, estymator taki dominuje klasyczny estymator (5.4), jeśli weźmiemy pod uwagę kwadratową funkcję straty.

Konstrukcja estymatora Rao-Blackwellizowanego polega na zauważeniu, że ciąg $X_1, X_2, \ldots$ symulowany przez algorytm MH powstaje dzięki wykorzystaniu *dwóch* próbek – $Y_1, Y_2, \ldots$ oraz $U_1, U_2, \ldots$, gdzie $Y_i \sim g(.|x_i)$, a $U_i \sim U$. Symbol U, zgodnie z wprowadzoną wcześniej notacją, oznacza tutaj rozkład jednostajny na przedziale jednostkowym, a zmienna U_i wykorzystywana jest do akceptacji lub odrzucenia punktu – kandydata Y_i zgodnie z prawdopodobieństwem akceptacji $p(x_i, Y_i)$.

Intuicyjnie, jeśli przy konstrukcji estymatora weźmiemy pod uwagę również kroki *odrzucone*, a nie tylko *zaakceptowane*, jak w klasycznym wzorze (5.4), zwiększy się ilość dostępnej nam informacji, przez co spadnie wariancja projektowanego estymatora. W celu wykorzystania powyższego wnioskowania możemy napisać

$$\hat{h}_f(X) = \frac{1}{n}\sum_{i=1}^{n} h(X_i) = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)\sum_{j=1}^{i} \mathbb{1}(X_j = Y_i) \ , \tag{6.11}$$

co prowadzi do wzoru wykorzystującego warunkową wartość oczekiwaną

$$\begin{aligned}\hat{h}_f(X) = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)\mathbb{E}\left(\sum_{j=1}^{i} \mathbb{1}(X_j = Y_i)|Y_1, Y_2, \ldots, Y_n\right) = \\ = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)\left(\sum_{j=1}^{i} \mathrm{P}(X_j = Y_i\,|Y_1, Y_2, \ldots, Y_n)\right) \ .\end{aligned} \tag{6.12}$$

Co istotne, możliwe jest bezpośrednie, numeryczne obliczenie prawdopodobieństw postaci $\mathrm{P}(X_i = Y\,|Y_1, Y_2, \ldots, Y_n)$ (patrz np. [6]). Dla przykładu, przy algorytmie MH wykorzystującym generowanie z niezależnej funkcji gęstości $g(y)$ (patrz (6.8) i (6.9)), wprowadźmy następujące oznaczenia

$$w_i = \frac{f(y_i)}{g(y_i)} \ , \ v_{ij} = \min\left(\frac{w_i}{w_j}, 1\right), \text{ dla } 0 \leqslant i < j \ , \tag{6.13}$$

$$z_{ii} = 1 \ , \ z_{ij} = \prod_{k=i+1}^{j} (1 - v_{ij}) \ , \text{ dla } i < j \ . \tag{6.14}$$

Jeśli dla ułatwienia założymy, że X_0 jest wygenerowane zgodnie z interesującą nas gęstością $f(.)$, to mamy wtedy następujące twierdzenie

Twierdzenie 6.5. *Estymator postaci*

$$\hat{h}_f(X) = \frac{1}{n+1}\sum_{i=0}^{n} \varphi_i h(Y_i) \tag{6.15}$$

jest estymatorem dla całki w problemie (5.2)*, gdzie*

$$\varphi_i = \tau_i \sum_{j=i}^{n} z_{ij} \tag{6.16}$$

i $\tau_i = \mathrm{P}(X_i = Y_i \,|Y_0, Y_1, \ldots, Y_n)$*, przy czym*

$$\tau_0 = 1 \ , \ \tau_i = \sum_{j=0}^{i-1} \tau_j z_{j(i-1)} v_{ji} \ \ dla\ i > 0 \ . \tag{6.17}$$

Twierdzenie to, kosztem dodatkowych obliczeń dla współczynników z_{ij}, umożliwia zmniejszenie wariancji estymatora (6.15) w porównaniu do (5.4). Co istotne, w (6.15) wykorzystujemy wszystkie kroki $Y_0, Y_1, \ldots$ wygenerowane przez algorytm MH, również te odrzucone w wyniku porównania z prawdopodobieństwem akceptacji $p(x, y)$.

6.2 Dwuwymiarowy próbnik Gibbsa

W celu lepszego zilustrowania działania próbnika Gibbsa rozpoczniemy od jego wersji dwuwymiarowej, aby potem przejść do wersji wielowymiarowej.

W przypadku dwuwymiarowym interesować nas będzie symulowanie wartości $(X_1, Y_1), (X_2, Y_2), \ldots$ wektora losowego (X, Y) o łącznej gęstości $f(x, y)$. Zakładamy ponadto, że potrafimy generować zmienne losowe dla obu gęstości warunkowych $f_{X|Y}(x|y)$ i $f_{Y|X}(y|x)$. Startując z dowolnie wybranej, początkowej wartości x_0, próbnik Gibbsa losuje kolejne wartości w cyklu następujących kroków:

Algorytm 6.6.

1. Wylosuj zmienną Y_i z gęstości $f_{Y|X}(.|x_{i-1})$
2. Wylosuj zmienną X_i z gęstości $f_{X|Y}(.|y_i)$

Jak łatwo zauważyć, próbnik Gibbsa w tej wersji polega na naprzemiennym generowaniu jednej ze zmiennych losowych X lub Y z odpowiedniej gęstości warunkowej pod warunkiem drugiej ze zmiennych. Innymi słowy, algorytm niejako „przeskakuje" pomiędzy „osiami" wektora losowego (X, Y).

Próbnik Gibbsa może być przede wszystkim zastosowany do losowania próbek z wektora losowego (X, Y) o trudnej do generowania gęstości łącznej $f(x, y)$. Co istotne, również podciągi $X_0, X_1, \ldots$ i $Y_0, Y_1, \ldots$ są ŁM, których rozkładami stacjonarnymi są odpowiednie gęstości brzegowe $f_X(.)$ i $f_Y(.)$. Umożliwia to zastosowanie próbnika Gibbsa również do otrzymywania próbek o odpowiednich gęstościach brzegowych, co jest przydatne np. w *modelach brakujących danych*. W tym przypadku, jeśli interesują nas zmienne o skomplikowanej gęstości $f_X(.)$, możliwe jest dodanie dodatkowej zmiennej Y i wykorzystywanie prostszych w generacji gęstości warunkowych $f_{X|Y}(x|y)$ i $f_{Y|X}(y|x)$. W takim modelu nie jest dla nas ważne symulowanie zmiennych o łącznej gęstości $f(x, y)$, ale otrzymanie próbek z odpowiedniej gęstości brzegowej. Zmienna Y pełni wtedy funkcję zmiennej pomocniczej (w j. ang. *auxiliary variable*), niebranej pod uwagę w wyniku, ale istotnej w procesie jego otrzymywania.

Podsumowanie powyższych uwag dotyczących zbieżności do odpowiednich gęstości dla próbnika Gibbsa znaleźć można w następującej definicji i twierdzeniu (patrz np. [30]).

Definicja 6.7. *Niech* $\mathbb{X} = \left(X^{(1)}, X^{(2)}, \ldots, X^{(m)}\right)$ *będzie* m*-wymiarowym wektorem losowym o gęstości łącznej* $f_{\mathbb{X}}\left(x^{(1)}, \ldots, x^{(m)}\right)$*, a* $f_{X^{(i)}}(.)$ *gęstością brzegową zmiennej* $x^{(i)}$*. Powiemy, że* $f_{\mathbb{X}}$ *spełnia warunek* dodatniości *(w j. ang.* positivity condition*), jeśli zachodzi następująca implikacja*

$$f_{X^{(i)}}\left(x^{(i)}\right) > 0 \text{ dla każdego } i = 1, \ldots, p \Rightarrow f_{\mathbb{X}}\left(x^{(1)}, \ldots, x^{(m)}\right) > 0 \ . \quad (6.18)$$

Innymi słowy, nośnik gęstości łącznej $f_{\mathbb{X}}$ jest iloczynem kartezjańskim nośników poszczególnych gęstości brzegowych.

Twierdzenie 6.8. *Niech łączna gęstość* $f(x, y)$ *spełnia warunek dodatniości, a jądro przejścia* $\mathcal{K}_{(X,Y)}((x_{i-1}, y_{i-1}), (x_i, y_i))$ *będzie funkcją absolutnie ciągłą. Wtedy ŁM generowany za pomocą próbnika Gibbsa jest nieprzywiedlny, powracający w sensie Harrisa i zbieżny do swego rozkładu stacjonarnego* $f(x, y)$*. Ponadto podciągi* $X_0, X_1, \ldots$ *i* $Y_0, Y_1, \ldots$ *są także ŁM zbieżnymi do swych odpowiednich rozkładów stacjonarnych* $f_X(.)$ *i* $f_Y(.)$*.*

6.3 Wielowymiarowy próbnik Gibbsa

W wielowymiarowym próbniku Gibbsa konstruowany jest ciąg m-wymiarowych wektorów losowych $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n, \ldots$ o gęstości łącznej $f(\mathbb{X})$. Przez $\left(X_i^{(1)}, \ldots, X_i^{(m)}\right)$ będziemy oznaczać wektor współrzędnych zmiennej losowej $\mathbb{X}_i$. W ogólnym przypadku każda z poszczególnych współrzędnych $X_i^{(j)}$ też może być wielowymiarowym wektorem losowym.

Przez $\mathbb{X}_i^{(-j)}$ oznaczymy wektor losowy $\left(X_i^{(1)}, \ldots, X_i^{(j-1)}, X_i^{(j+1)}, \ldots, X_i^{(m)}\right)$, czyli wektor pozbawiony swojej j-współrzędnej. Ponadto założymy, iż możliwe jest efektywne generowanie zmiennych losowych z poszczególnych gęstości warunkowych $f_{X^{(j)}|\mathbb{X}^{(-j)}}\left(.|x^{(1)}, \ldots, x^{(j-1)}, x^{(j+1)}, \ldots, x^{(m)}\right)$ dla $j = 1, \ldots, m$. Gęstości takie określane są jako *pełne warunkowe* (w j. ang. *full conditionals*). Jak widzimy, jest to odpowiednik podobnego warunku dla dwuwymiarowego próbnika Gibbsa.

Dowolną wartość startową $\left(x_0^{(1)}, \ldots, x_0^{(m)}\right)$ dla algorytmu oznaczymy przez $\boldsymbol{x}_0$. Wtedy, przy podanych powyżej założeniach, próbnik Gibbsa generuje kolejne elementy ciągu według następującego algorytmu:

Algorytm 6.9.

1. Generuj $X_{i+1}^{(1)}$ z gęstości $f_{X^{(1)}|\mathbb{X}^{(-1)}}\left(.|x_i^{(2)}, \ldots, x_i^{(m)}\right)$
2. Generuj $X_{i+1}^{(2)}$ z gęstości $f_{X^{(2)}|\mathbb{X}^{(-2)}}\left(.|x_{i+1}^{(1)}, x_i^{(3)}, \ldots, x_i^{(m)}\right)$
3. Generuj $X_{i+1}^{(3)}$ z gęstości $f_{X^{(3)}|\mathbb{X}^{(-3)}}\left(.|x_{i+1}^{(1)}, x_{i+1}^{(2)}, x_i^{(4)}, \ldots, x_i^{(m)}\right)$
4. ...

m. Generuj $X_{i+1}^{(m)}$ z gęstości $f_{X^{(m)}|\mathbb{X}^{(-m)}}\left(.|x_{i+1}^{(1)}, x_{i+1}^{(2)}, \ldots, x_{i+1}^{(m-1)}\right)$

Powyższy algorytm jest bezpośrednim przeniesieniem przypadku dwuwymiarowego w m-wymiarowy. Również tutaj następuje generowanie poszczególnych zmiennych z kolejnych gęstości warunkowych, co odpowiada „poruszaniu się" po kolejnych „osiach" wektora losowego $\mathbb{X}$.

Istnieją rozmaite modyfikacje powyższego algorytmu. Jedna z nich polega na zastosowaniu permutacji zamiast systematycznej sekwencji „ruchów" względem kolejnych współrzędnych wektora losowego $\mathbb{X}$. W algorytmie tym najpierw losowana jest dowolna permutacja Σ zbioru m-elementowego, a następnie wykorzystuje się tą permutację do wyboru kolejności generowania względem poszczególnych gęstości warunkowych.

Inna modyfikacja polega na wykorzystaniu dodatkowego, niezależnego rozkładu pomocniczego V o nośniku złożonym z m elementów (patrz np. [24]). Przy takiej modyfikacji, najpierw generuje się zmienną losową J zgodnie z rozkładem V, aby dla wygenerowanej wartości j z tego rozkładu, dokonać losowania nowej wartości współrzędnej $X_{i+1}^{(j)}$ z gęstości $f_{X^{(j)}|\mathbb{X}^{(-j)}}\left(.|\boldsymbol{x}^{(-j)}\right)$. W tej modyfikacji dokonujemy więc losowego wyboru „osi", względem której następuje „ruch" próbnika Gibbsa.

Podobnie jak w dwuwymiarowym przypadku, dla wielowymiarowego próbnika Gibbsa istnieją twierdzenia dotyczące zbieżności ŁM $\mathbb{X}_1, \mathbb{X}_2, \ldots, \mathbb{X}_n, \ldots$ do gęstości łącznej $f(\mathbb{X})$. W oczywisty sposób, ich założenia powinny być łatwe do sprawdzenia w praktycznych zastosowaniach. Podamy teraz przykład jednego z takich twierdzeń.

Twierdzenie 6.10. *Łańcuch Markowa wygenerowany próbnikiem Gibbsa jest nieprzywiedlny i nieokresowy, jeśli gęstości warunkowe $f_{X^{(j)}|\mathbb{X}^{(-j)}}$ spełniają następujące warunki:*

1. *Niech $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(m)})$ i $\boldsymbol{x}' = (x'^{(1)}, \ldots, x'^{(m)})$ oraz istnieje $\delta > 0$ dla której $\boldsymbol{x}, \boldsymbol{x}' \in \mathrm{supp}(f_{\mathbb{X}})$, $|\boldsymbol{x} - \boldsymbol{x}'| < \delta$ i*

$$f_{X^{(j)}|\mathbb{X}^{(-j)}}\left(x^{(j)} \middle| x^{(1)}, \ldots, x^{(j-1)}, x'^{(j+1)}, \ldots, x'^{(m)}\right) > 0 \text{ dla } i = 1, \ldots, m \tag{6.19}$$

2. *Istnieje $\delta' < \delta$, taka, że każda para $\boldsymbol{x}, \boldsymbol{x}' \in \mathrm{supp}(f_{\mathbb{X}})$ może być połączona skończoną ilością kul o promieniu δ' mających parami niepuste przecięcia.*

Mimo skomplikowanego sformułowania powyższego twierdzenia, jego zasadniczym elementem są dwa warunki. Pierwszy z nich określa, iż istnieje niezerowe prawdopodobieństwo przejścia pomiędzy dwoma dowolnymi, dostatecznie bliskim sobie stanami $\boldsymbol{x}$ i $\boldsymbol{x}'$. Celem drugiego jest wymaganie, aby wielowymiarowy próbnik Gibbsa poruszał się po spójnej przestrzeni stanów, tzn. miał nośnik spójny w sensie topologicznym.

6.4 Algorytm MH a próbnik Gibbsa

Obecnie skupimy się na podobieństwach i różnicach występujących pomiędzy dwoma omówionymi wcześniej algorytmami generowania ŁM w metodach MCMC – algorytmie Metropolisa – Hastingsa i próbniku Gibbsa.

Rozpoczniemy od istotnego teoretycznie twierdzenia, które niejako uprawnia do tego typu porównania.

Twierdzenie 6.11. *Wielowymiarowy próbnik Gibbsa jest równoważny połączeniu ze sobą m algorytmów MH o prawdopodobieństwach akceptacji równych zawsze jeden.*

Liczba m występująca w powyższym twierdzeniu jest, jak pamiętamy, wymiarem wektora $\mathbb{X}$ generowanym przez próbnik Gibbsa (patrz rozdział 6.3). Dowód tego twierdzenia znaleźć można w [30].

Mimo stwierdzonej powyżej *teoretycznej* odpowiedniości pomiędzy oboma algorytmami, istnieją pewne istotne różnice w ich *praktycznym* zastosowaniu. Przede wszystkim próbnik Gibbsa wymaga dość istotnej wiedzy o postaci funkcji łącznej $f(\mathbb{X})$, z której chcemy uzyskać próbki. Wiedza ta to przynajmniej znajomość poszczególnych pełnych warunkowych funkcji gęstości $f_{X^{(j)}|\mathbb{X}^{(-j)}}\left(.|\boldsymbol{x}^{(-j)}\right)$ wraz z umiejętnością symulowania wartości zmiennych losowych dla tych gęstości. Jednocześnie konieczność stosowania pełnych gęstości warunkowych powoduje brak takiej swobody w konstrukcji algorytmu, jaką zapewniała możliwość wyboru prawie dowolnej gęstości instrumentalnej w algorytmie MH. Może to być rozpatrywane zarówno jako wada, jak i zaleta próbnika Gibbsa – z jednej strony trudniejsza jest jego optymalizacja i mniejsza elastyczność, z drugiej – trudniej o nieodpowiedni wybór, np. błędnej lub zdegenerowanej lokalnie gęstości.

Po drugie, z samej swej natury, próbnik Gibbsa jest niejako „wielowymiarowy" – sama jego konstrukcja wymaga stosowania minimum dwuwymiarowych zmiennych losowych (patrz rozdział 6.2). Nawet przy modelach brakujących danych, niezbędne jest zastosowanie drugiej losowej zmiennej pomocniczej.

Inna różnica pomiędzy algorytmami wynika bezpośrednio z twierdzenia 6.11 – w przypadku próbnika Gibbsa każda z proponowanych wartości jest *zawsze* akceptowana, gdy dla algorytmu MH zdarzają się odrzucenia nowych wartości na korzyść starych z prawdopodobieństwem $1 - p(x_{i-1}, Y_i)$ (patrz (6.1)).

W literaturze rozważane są rozmaite *połączenia* obu typów algorytmów, zwane *hybrydowymi algorytmami MCMC* (w j. ang. *hybrid MCMC algorithms*). Połączenia te mogą mieć postać *mieszanin* lub *cykli* (patrz np. [34]), zgodnie z poniższą definicją.

Definicja 6.12. *Niech $\mathcal{K}_X^{(1)}, \mathcal{K}_X^{(2)}, \ldots, \mathcal{K}_X^{(p)}$ będą jądrami przejścia dla kroków pewnego algorytmu generującego ŁM. Wtedy* mieszaniną *nazwiemy algorytm związany z jądrem*

$$\mathcal{K}_X = a_1\mathcal{K}_X^{(1)} + \ldots + a_p\mathcal{K}_X^{(p)} \ , \tag{6.20}$$

gdzie $a_1, \ldots, a_p$ jest pewnym rozkładem prawdopodobieństwa. Z kolei cyklem *nazwiemy zaś algorytm o jądrze przejścia zadanym złożeniem funkcji*

$$\mathcal{K}_X = \mathcal{K}_X^{(1)} \circ \ldots \circ \mathcal{K}_X^{(p)} \ . \tag{6.21}$$

Innymi słowy, hybrydowy algorytm MCMC polega na stworzeniu algorytmu złożonego jednocześnie z odpowiednio dobranych algorytmów MH i próbnika Gibbsa. Przykładowo, algorytm hybrydowy składać się może głównie z kroków próbnika Gibbsa, ale co ustalony wcześniej p-ty krok wykorzystywany jest algorytm MH – mamy wtedy do czynienia z cyklem w myśl definicji 6.12. Innym przykładem może być wykorzystywanie algorytmu MH z pewnym ustalonym wcześniej prawdopodobieństwem wśród kroków próbnika Gibbsa.

6.5 Przykładowe zastosowanie metody MCMC

Zastosowania symulacyjnych metod MCMC, jak zostało to wspomniane na początku tego rozdziału, obejmują bardzo szeroki zakres dziedzin i aplikacji praktycznych. W związku z tym omówimy tutaj pokrótce tylko jeden przykład zastosowania, związanego z odszumianiem i analizowaniem obrazów cyfrowych (patrz np. [21, 22]).

Konieczność wykorzystywania metod MCMC przy odszumianiu i analizowaniu zdjęć związana jest nie tylko z częstokroć bardzo dużym rozmiarem przestrzeni stanów, wynikającym z rozdzielczości zdjęcia, np. 2048^2, ale i z wprowadzeniem ewentualnego trzeciego wymiaru, czyli np. czasu w filmach. Genezy stosowanych przy tym algorytmów należy upatrywać w zastosowaniach fizycznych, np. w badaniach nad magnetycznością ciał stałych (patrz np. [2]).

W niniejszym rozdziale przyjrzymy się przede wszystkim problemowi *odszumiania obrazów cyfrowych*. Przez $\mathbb{X}$ oznaczymy prawdziwy obraz, składający się z pikseli, czyli z pojedynczych elementów obrazu $x^{(i)}$. Zazwyczaj piksele $x^{(i)}$ tworzą ściśle określoną dwuwymiarową strukturę, np. prostokątną tablicę w przypadku standardowych zdjęć, a ich wartości mogą być binarne (np. dla zdjęć czarno-białych), dyskretne (w przypadku zdjęć w odcieniach szarości) lub rzeczywiste. Wartości przyjmowane przez piksele nazywać będziemy dalej *kolorami*.

Ogólniej rzecz ujmując, obraz $\mathbb{X}$ jest elementem przestrzeni $\mathcal{A}^{\mathcal{S}}$, gdzie $\mathcal{A}$ jest zbiorem dopuszczalnych kolorów, a $\mathcal{S}$ – zbiorem wszystkich pikseli w obrazie. Na zbiorze $\mathcal{S}$ zadany jest graf K zadający strukturę sąsiedztw dla poszczególnych pikseli. Dla zdjęć przykładem takiego zbioru sąsiedztw pojedynczego piksela $x^{(i)}$ mogą być wszystkie piksele stykające się brzegiem lub wierzchołkiem z tym pikselem.

Obserwowany przez nas, zarejestrowany obraz $\mathbb{Y}$ składa się z kolei z pikseli oznaczanych przez $y^{(i)}$. Podobnie jak poprzednio, $\mathbb{Y} \in \mathcal{A}^{\mathcal{S}}$, choć w ogólności przestrzeń kolorów i zbiór pikseli dla obrazu zarejestrowanego mogą być inne niż dla obrazu rzeczywistego $\mathbb{X}$.

Zakładamy, że zarejestrowany obraz $\mathbb{Y}$ jest zaszumionym zniekształceniem obrazu rzeczywistego $\mathbb{X}$ zgodnie z ogólnym modelem szumu zadanym przez prawdopodobieństwo

$$\mathrm{P}\left(\mathbb{Y}|\mathbb{X}\right) = \prod_{i\in\mathcal{S}} \mathrm{P}\left(y^{(i)} \left| x^{(\nu_i)} \right.\right) \ , \tag{6.22}$$

gdzie ν_i jest zbiorem wszystkich pikseli, które wpływają na stan piksela $y^{(i)}$. Jak łatwo zauważyć, w celu pełnego zdefiniowania modelu zaszumienia (6.22), niezbędne jest określenie postaci funkcji $\mathrm{P}\left(.|.\right)$, np. jako szumu gaussowskiego z bezpośrednim oddziaływaniem wartości piksela $x^{(i)}$ na $y^{(i)}$

$$\mathrm{P}\left(y^{(i)} \left| x^{(i)} \right.\right) \sim N(x^{(i)}; \sigma^2) \ , \tag{6.23}$$

gdzie σ^2 jest ustaloną wariancją – poziomem szumu tła. Dla obrazów o binarnych kolorach szum może być generowany poprzez losową zamianę „czarnego" w „białe" i odwrotnie, zgodnie z regułą

$$\mathrm{P}\left(y^{(i)} \left| x^{(i)} \right.\right) = \begin{cases} y^{(i)} = x^{(i)} & \text{z prawd. } p \\ y^{(i)} \neq x^{(i)} & \text{z prawd. } 1-p \end{cases} \ , \tag{6.24}$$

gdzie $1-p$ jest prawdopodobieństwem zmiany koloru piksela.

Pełne wyspecyfikowanie modelu wymaga jeszcze poczynienia założeń odnośnie prawdopodobieństw *a priori* dla pikseli rzeczywistego obrazu $\mathbb{X}$. Założenia te mogą mieć postać bardzo ogólną, jak i szczegółową, jeśli dysponujemy dodatkowymi informacjami na temat odszumianego obrazu. Przykładem takiego ogólnego założenia jest *model Pottsa*, wykorzystywany także w zagadnieniach fizycznych

$$\mathrm{P}(\mathbb{X}) \propto \prod_{i \in \mathcal{S}} \exp\left(-\beta \sum_{i \sim j} \mathbb{1}\left(x^{(i)} = x^{(j)}\right)\right) \ , \tag{6.25}$$

gdzie $i \sim j$ oznacza wszystkie piksele znajdujące się w pewnym określonym *sąsiedztwie* piksela $x^{(i)}$, zgodnie z zadanym wcześniej grafem K. Parametr β ma wpływ na wagę informacji *a priori* względem informacji obserwowanej. Wzór (6.25) oznacza intuicyjne przekonanie, że sąsiednie piksele, np. bezpośrednio stykające się ze sobą, mają ten sam kolor. Sprzyja to powstawaniu dużych, jednokolorowych obszarów.

Ze wzoru Bayesa mamy prawdopodobieństwo *a posteriori*

$$\mathrm{P}\left(\mathbb{X}|\mathbb{Y}\right) \propto \mathrm{P}(\mathbb{X})\,\mathrm{P}\left(\mathbb{Y}|\mathbb{X}\right) \ , \tag{6.26}$$

skąd, dla modelu zadanego przez (6.22) i (6.25), łatwo otrzymujemy wzór na pełne warunkowe prawdopodobieństwa (porównaj z definicją wprowadzoną w rozdziale 6.3)

$$\begin{aligned} \mathrm{P}\left(x^{(i)} \,\middle|\, \boldsymbol{x}^{(-i)}, \mathbb{Y}\right) \propto \\ \propto \exp\left(-\beta \sum_{i \sim j, i \neq j} \mathbb{1}\left(x^{(i)} = x^{(j)}\right)\right) \prod_{j : j \in \nu_i} \mathrm{P}\left(y^{(j)} \,\middle|\, x^{(\nu_i)}\right) \ . \end{aligned} \tag{6.27}$$

Zauważmy, że z założeń modelu wynika, iż skoro interesuje nas postać obrazu rzeczywistego $\mathbb{X}$, to oznacza, że szukamy *najbardziej prawdopodobnej* wartości rozkładu. Dysponując pełnymi warunkowymi prawdopodobieństwami (6.27), możemy do rozwiązania tego problemu bezpośrednio skorzystać z wielowymiarowego próbnik Gibbsa (patrz rozdział 6.3). Na przykład, jeśli przez sąsiedztwo piksela i-tego zdefiniujemy jego ośmiu najbliższych sąsiadów (cztery piksele sąsiadujące krawędziami i cztery wierzchołkami), a jako model szumu wykorzystamy (6.24), to otrzymujemy

$$\begin{aligned} \mathrm{P}\left(x^{(i)} \,\middle|\, \boldsymbol{x}^{(-i)}, \mathbb{Y}\right) \propto \\ \propto \exp\left(-\beta \sum_{i \sim j, i \neq j} \mathbb{1}\left(x^{(i)} = x^{(j)}\right) + \ln p\ \mathbb{1}\left(x^{(i)} = x^{(j)}\right) + \right. \\ \left. + \ln(1-p)\ \mathbb{1}\left(x^{(i)} \neq x^{(j)}\right)\right) \ . \end{aligned} \tag{6.28}$$

Jak łatwo zauważyć, ponieważ sąsiedztwo obejmuje w tym momencie tylko osiem innych pikseli, obliczenie czynnika normującego dla (6.28) jest bardzo łatwe. W konsekwencji umożliwia to bezproblemowe wykorzystanie próbnika Gibbsa.

Model Pottsa opisany wzorem (6.25) najlepiej odpowiada przypadkowi czarno-białych zdjęć, czyli binarności stanów pikseli obrazu. Dla skali szarości lub obrazów kolorowych lepszym wyborem jest założenie prawdopodobieństwa *a priori* danego przez

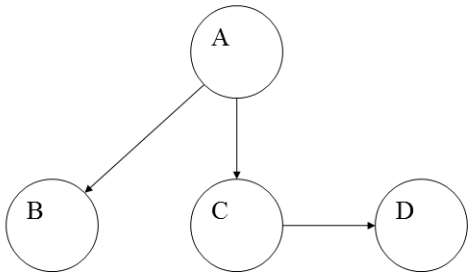
$$\mathrm{P}(\mathbb{X}) \propto \prod_{i \in \mathcal{S}} \exp\left(-\beta \sum_{i \sim j} \psi\left(x^{(i)} - x^{(j)} \right) \right) , \tag{6.29}$$

gdzie $\psi(.)$ jest pewną symetryczną funkcją (patrz np. [1, 13]). W tego typu problemach pełne warunkowe prawdopodobieństwa mogą mieć zbyt skomplikowaną postać do bezpośredniego wykorzystania próbnika Gibbsa i niezbędne może być użycie algorytmu MH (patrz rozdział 6.1).

Innym przykładem związanym z analizowaniem i odszumianiem obrazów cyfrowych jest wykorzystanie metod MCMC w obróbce obrazów otrzymanych w tomografii komputerowej SPECT i PET (patrz np. [15, 21]). W modelach tych $\mathbb{X}$ jest opisem fizycznego modelu – odwzorowaniem ciała pacjenta poddanego badaniu tomograficznemu. Do ciała pacjenta wprowadzany jest radioaktywny izotop, emitujący pozytrony w wyniku rozpadu jąder atomów. Pozytrony te anihilując z elektronami, powodują powstanie fotonów, rejestrowanych następnie przez aparaturę medyczną. W takim ujęciu $\mathbb{Y}$ nie jest jedynie zaszumionym zniekształceniem obrazu rzeczywistego $\mathbb{X}$, ale zupełnie inną przestrzenią – modelem odwzorowującym sensory rejestrujące fotony. Co istotne, w tego typu zastosowaniu możliwe jest wykorzystanie przedstawionej powyżej metody odszumiania obrazów, ale w celu ich rekonstrukcji. Najogólniej rzecz ujmując, w tym przypadku problemem do rozwiązania jest znalezienie miejsca i wielkości emisji pozytronów, czyli odtworzenie modelu nieznanej emisji $\mathbb{X}$ bazując na zaobserwowanych wynikach sensorów $\mathbb{Y}$. Dodatkowo sam proces odwzorowywania emisji przez aparaturę skutkuje dodatkowymi zmianami w stosunku do omówionej wcześniej metody odszumiana - np. sąsiedztwo „piksela" $x^{(i)}$ nie musi być zbiorem zwartym. Przy zastosowaniu metod MCMC możliwe jest jednak pozytywne rozstrzygnięcie, z którego miejsca ciała pacjenta emitowane są rejestrowane przez aparaturę fotony.

W tym miejscu warto jeszcze wspomnieć o modelach bayesowskich. Zwróćmy bowiem uwagę, iż opisana powyżej metoda odszumiania obrazu odpowiada dokładnie zasadom wnioskowania bayesowskiego. W szczególności próbnik Gibbsa jest znakomicie dostosowany do tego typu zadań, dzięki wykorzystywaniu pełnych warunkowych gęstości przy generowaniu zmiennych z poszukiwanego rozkładu.

Zalety próbnika Gibbsa są szczególnie widoczne przy analizowaniu złożonych, wielopoziomowych modeli bayesowskich opisywanych zazwyczaj graficznie za pomocą DAG-ów, tzn. skierowanych, niecyklicznych grafów (w j. ang. *direct, acyclic graph*). DAG-i składają się z węzłów połączonych strzałkami. *Rodzicami* danego węzła w grafie DAG nazwiemy wszystkie węzły bezpośrednio połączone z rozpatrywanym węzłem, których strzałki skierowane są *w kierunku do* tego węzła, a *dziećmi* – węzły, których strzałki skierowane są *od* rozpatrywanego węzła. Dla przykładu – na rys. 6.1, dziećmi węzła `A` są węzły `B` i `C`, rodzicem węzła `C` jest węzeł `A`, a dzieckiem `C` – węzeł `D`. W zastosowaniach statystycznych węzły reprezentują rozkłady poszczególnych parametrów modelu, od których to zależą rozkłady parametrów w węzłach – dzieciach.



Rysunek 6.1: Rodzice i dzieci w DAG-ach

W tego typu modelach wykorzystanie metod MCMC jest możliwe dzięki pewnej obserwacji. Otóż rozkład łączny wszystkich zmiennych można przedstawić jako iloczyn rozkładów warunkowych poszczególnych składowych, przy czym rozkłady te zależą tylko od parametrów węzłów – rodziców (patrz np. [23]), tzn.

$$\mathrm{P}(\mathbb{V}) = \prod_{v \in \mathbb{V}} \mathrm{P}\left(v \mid \text{ rodzice } v\right) \ , \tag{6.30}$$

gdzie $\mathbb{V}$ są wszystkimi zmiennymi losowymi w rozpatrywanym modelu. Umożliwia to szybkie konstruowanie niezbędnych w próbniku Gibbsa pełnych warunkowych rozkładów prawdopodobieństw (patrz np. [15]).

6.6 Zalety i wady metod MCMC

Jak zostało to podkreślone w poprzednich rozdziałach, największą zaletą metod MCMC w porównaniu do metod MC jest brak konieczności bezpośredniego generowania próbek z pewnej, być może bardzo skomplikowanej, funkcji gęstości $f(x)$ np. w problemie (5.2). Zamiast tego stosujemy dość dowolnie wybraną gęstość instrumentalną $g(y|x)$ – w algorytmie MH (patrz rozdział 6.1), albo pełne gęstości warunkowe $f_{X^{(j)}|\mathbb{X}^{(-j)}}\left(.|\boldsymbol{x}^{(-j)}\right)$ – w próbniku Gibbsa (patrz rozdział 6.3). Pozwala to w łatwy sposób ominąć niektóre problemy zasygnalizowane w rozdziale 5.3.

Możliwość skorzystania z odpowiednich twierdzeń ergodycznych, które zapewniają zbieżność wygenerowanego ŁM do gęstości stacjonarnej $f(x)$ jest jednocześnie największą ze słabości metod MCMC. Zauważmy bowiem, że na stworzony odpowiednim algorytmem ŁM wpływ mają jednocześnie punkt startowy X_0 oraz liczba wykonanych przez ten algorytm kroków n. W oczywisty sposób, ŁM powinien szybko „zapomnieć" o wybranej lub wylosowanej przez użytkownika wartości startowej, aby nie miała ona wpływu na końcowy estymator i nie powodowała jego statystycznego obciążenia. Z drugiej strony, niezbędna jest również pewna minimalna liczba kroków algorytmu, aby estymator zbliżył się do poszukiwanej przy jego pomocy wartości.

Nie jest przy tym możliwe, jak w przypadku metod MC, bezpośrednie skorzystanie z centralnych twierdzeń granicznych gwarantujących stworzenie przedziału ufności dla estymatora na z góry określonym poziomie ufności (patrz

np. twierdzenie 5.2). Niestety, aby efektywnie wykorzystać odpowiedniki CTG dla ŁM, wymagane jest posiadania estymatorów *kowariancji* zmiennych losowych dla poszczególnych kroków ŁM (patrz założenia twierdzeń 1.72 i 1.74). W oczywisty sposób estymowanie kowariancji dla zmiennych nie posiadających jednakowego rozkładu jest zadaniem szczególnie trudnym i wymagającym zastosowania dodatkowych symulacji.

W związku z powyższym, istotnym zagadnieniem w metodach MCMC jest problem *diagnostyki*, tzn. określania takich warunków i ograniczeń dla algorytmu przy których tworzony metodą MCMC estymator „w odpowiedni sposób" estymuje pożądaną przez nas wartość. Zagadnieniem tym szczegółowo zajmiemy się w rozdziale 6.7.

6.7 Diagnostyka metod MCMC

Jak zostało to zasygnalizowane w rozdziale 6.6, jednym z najistotniejszych dla praktyki metod MCMC problemów jest *zagadnienie diagnostyki*. W różnych pracach istnieją różne definicje tego zagadnienia i odmienne podejścia do jego rozwiązywania. W niniejszej przeglądzie wzorować się będziemy głównie na omówieniu przedstawionym w [30], jako jednym z najbardziej przekrojowych i jednocześnie najlepiej systematyzujących różne źródła.

Najgólniej rzecz ujmując, diagnostyka metod MCMC polega na określeniu *ile początkowych wartości wygenerowanego ŁM należy odrzucić* oraz *w którym momencie można symulację zakończyć*. Należy podkreślić, że wielu autorów zajmuje się w swoich pracach *jedynie* drugim z przedstawionych pytań, uznając je za znacznie istotniejsze i deklaratywnie przyjmując odrzucenie np. 10% pierwszych kroków zasymulowanego łańcucha $X_0, X_1, \ldots, X_n$. Oczywiście, tego typu podejście może prowadzić do statystycznego obciążenia konstruowanego estymatora, związanego ze zbytnim wpływem wartości początkowej na uzyskiwany wynik.

Innym problemem, wartym zauważenia, jest fakt kłopotów z porównywalnością jakości rozmaitych metod diagnozy. Bezpośrednio idzie za tym niejasność sformułowania i chyba także niewielka celowość pytania, która z rozważanych dalej metod diagnostycznych jest *najlepsza*. Z jednej strony trudnym jest bowiem porównywanie metod heurystycznych, bardzo powszechnych w diagnostyce MCMC, z metodami bardziej teoretycznymi, korzystającymi ze skomplikowanych nierzadko twierdzeń. Wynika to zarówno z subiektywności ocen wysuwanych na podstawie intuicji i wiedzy badacza, obserwującego np. wykres zbieżności estymatora, z dość jednak obiektywnymi miarami bazującymi na wariancji estymatora. Co więcej, metody heurystyczne częstokroć wymagają znacznie mniej *wiedzy* na temat teoretycznych własności tworzonego ŁM i samego algorytmu, niewykorzystując jego specyficznych w danym przypadku właściwości. W oczywisty sposób trudno jest także porównywać różne metody *stricte* teoretyczne, ale bazujące na *odmiennych*, szczególnych własnościach ŁM, takich jak np. posiadanie atomów w przestrzeni stanów. Nie jest bowiem możliwe porównanie metod, które z konieczności odnoszą się do zupełnie różnych typów ŁM.

Drugim z elementów, które wpływają na kwestię możliwości porównywania metod diagnozy jest wspomniana wcześniej praktyka autorów, zwracających uwagę na *odmienne* aspekty samego zagadnienia diagnozy. Łatwo zauważyć, że

jeśli dwie odmienne metody biorą pod uwagę zupełnie inne *kryteria zbieżności* w diagnostyce, wynik ich jednoczesnego zastosowania może też być zupełnie różny – ŁM dobrze estymujący szukaną wielkość według jednej metody, może wymagać większej, dodatkowej liczby kroków, jeśli weźmiemy pod uwagę inne kryterium diagnostyczne.

W związku z tym, wydaje się, iż wszelkie metody diagnostyczne należy rozpatrywać raczej jako *zestaw* wzajemnie się uzupełniających narzędzi, które należy stosować *wspólnie* dla rozstrzygania problemu zbieżności ŁM. Brak zbieżności wykryty za pomocą jednej metody, może wszak pozostać niezauważony dla pozostałych kryteriów diagnostycznych. Dopiero zastosowanie dużej liczby różnych metod, które będą zgodnie stwierdzać nastąpienie zbieżności, może *przybliżyć* nas do postawienia prawidłowej diagnozy.

6.8 Kryteria zbieżności w diagnostyce

W niniejszym rozdziale pokrótce przyjrzymy się kilku metodom diagnostycznym, podzielonym, zgodnie z podejściem zaprezentowanym w [30], według różnych *kryteriów zbieżności*.

6.8.1 Zbieżność do rozkładu stacjonarnego

W praktyce metod MCMC każdy generowany przez odpowiedni algorytm ŁM $X_0, X_1, \ldots$ startuje z pewnej, uprzednio wybranej lub wylosowanej, wartości początkowej X_0. Jak zostało wspomniane wcześniej, wartość startowa nie powinna mieć wpływu na uzyskiwany dzięki metodzie MCMC estymator, tzn. nie może prowadzić do jego obciążenia lub zafałszowania. Jeśli przez $\mathrm{P}^n_{x_0}$ oznaczymy rozkład prawdopodobieństwa dla n-tego kroku ŁM, a przez π_X rozkład stacjonarny tego ŁM, to interesować się będziemy *odległością* pomiędzy $\mathrm{P}^n_{x_0}(.)$ a $\pi_X(.)$. Odległość należy tutaj rozumieć w sensie pewnej wybranej normy dla rozkładów prawdopodobieństwa obliczanej względem stanów z przestrzeni stanów łańcucha $\mathcal{X}$. Przykładem takiej odległości może być chociażby norma *supremum*:

$$\left\|\mathrm{P}^n_{x_0}(.) - \pi_X(.)\right\|_{\sup} = \sup_{x\in\mathcal{X}} \left|\mathrm{P}^n_{x_0}(x) - \pi_X(x)\right| \ . \tag{6.31}$$

Jedną z prostych metod diagnostycznych mających zastosowanie w tym zagadnieniu, jest wykorzystanie modyfikacji dowolnego nieparametrycznego testu zgodności, np. testu Kołmogorowa-Smirnowa. Z samej definicji wynika bowiem, że jeśli wyraz X_l ŁM pochodzi z rozkładu stacjonarnego π_X, to także zmienna X_{l+1} będzie mieć taki sam rozkład.

W celu wykorzystania opisywanej metody dokonujemy podziału ŁM na dwie *oddzielne* połowy – $X_1, \ldots, X_{n/2}$ oraz $X_{n/2+1}, \ldots, X_n$. Oczywiście bezpośrednie zastosowanie odpowiednika statystyki testowej Kołmogorowa-Smirnowa nie jest możliwe ze względu na *własność Markowa* łańcucha, czyli istnienie korelacji pomiędzy kolejnymi jego wyrazami. To niekorzystne zjawisko można próbować wyeliminować dokonując tzw. *subsamplingu*, tzn. wykorzystując tylko co k-wyraz w obu podłańcuchach z ustalonym krokiem k. Otrzymujemy dzięki temu podciągi postaci $X_1, X_{k+1}, \ldots$ – oznaczany dalej jako $V_1, V_2, \ldots$, oraz podciąg $X_{n/2+1}, X_{n/2+1+k}, \ldots$ – oznaczany przez $V_1^{'}, V_2^{'}, \ldots$. Choć sposób ten zmniej-

sza oczywiście stopień zależności występujący pomiędzy próbkami, nie eliminuje całkowicie tej komplikacji, a ponadto prowadzi do częściowej straty informacji.

Następnym krokiem jest zastosowanie zmodyfikowanej statystyki Kołmogorowa-Smirnowa

$$T_{\mathrm{KS}} = \frac{1}{M} \sup_x \left| \sum_{i=1}^{M} \Big(\mathbb{1}(V_i \in (0,x)) - \mathbb{1}(V_i^{'} \in (0,x)) \Big) \right| , \tag{6.32}$$

gdzie M jest największą liczbą podzielną przez k i mniejszą niż $n/2$, czyli maksymalnym numerem wyrazu w podciągach V_i i $V_i^{'}$.

Możliwe jest przy tym otrzymanie pełnej postaci gęstości dla statystyki (6.32), ale niestety tylko asymptotycznie w przypadku $M \to \infty$. Statystyka $\sqrt{M}T_{\mathrm{KS}}$ przy założeniu prawdziwości hipotezy zerowej, tzn. stacjonarności ŁM, ma wtedy dystrybuantę postaci (patrz np. [30])

$$D_{\sqrt{M}T_{\mathrm{KS}}}(x) = 1 - \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2x^2} \ . \tag{6.33}$$

Dokładne wyliczenie wartości krytycznej statystyki (6.32) dla dowolnego M skończonego jest niestety znacznie utrudnione ze względu na korelacje występujące pomiędzy poszczególnymi elementami ŁM. W dodatku trudność problemu wzrasta przy zastosowaniu metody *subsamplingu* z powodu wprowadzanego przez nią czynnika losowego, który skutkuje skomplikowaniem postaci zależności pomiędzy zmiennymi.

Inną metodą jest próba zbadania, czy ŁM symulowany metodą MCMC $X_0, X_1, \ldots$ „odwiedził" wszystkie stany swej dziedziny $\mathcal{X}$. Jeśli tak jest, to estymator całki

$$\int_{\mathcal{X}} f(x)dx \tag{6.34}$$

powinien być oczywiście zbieżny do jedynki. W celu sprawdzenia tej zbieżności w literaturze zaproponowano estymację (6.34) poprzez sumy Riemanna postaci

$$\sum_{k=0}^{n-1} (X_{k+1} - X_k)\, f(X_k) \ . \tag{6.35}$$

W przypadku gdy przestrzeń $\mathcal{X}$ jest jednowymiarowa, (6.35) powinien zbiegać do jedynki nawet, gdy zmienne X_k nie są generowane dokładnie z gęstości $f(x)$ (patrz np. [30]).

6.8.2 Zbieżność do średniej

Bardzo istotnym wskaźnikiem zbieżności dla metody MCMC jest kwestia *odległości* pomiędzy *estymatorem* zbudowanym w oparciu o próbkę z trajektorii $X_k, X_{k+1}, \ldots$ a poszukiwaną przez nas *średnią* $\mathbb{E}_f h(X)$ w problemie (5.2). Podobnie jak przy badaniu zbieżności do rozkładu stacjonarnego, *odległość* należy rozumieć w sensie pewnej normy lub jako minimalizację *błędu estymacji* poprzez zmniejszenie wariancji estymatora.

Odpowiedź na powyższe pytanie możemy traktować jako *drugi etap* w diagnozowaniu zbieżności, po stwierdzeniu odpowiednio małej odległości pomiędzy

rozkładami $\mathrm{P}^n_{x_0}(.)$ a $\pi_X(.)$. Dlatego też, zgodnie z tym, co napisaliśmy wcześniej, nasz estymator budujemy na podstawie ŁM o odrzuconych pierwszych k krokach. Jednak w celu ułatwienia notacji w dalszej części tego rozdziału używać będziemy prostszego zapisu $X_0, X_1, \ldots$, pamiętając przy tym o konieczności usunięcia pewnej ilości początkowych wartości ŁM.

Zauważmy, że właśnie kwestia zbieżności do średniej $\mathbb{E}_f h(X)$ jest jedną z najistotniejszych kwestii dla metod MCMC – bardzo wiele praktycznych problemów można zapisać właśnie jako rozwiązanie problemu (5.2).

Zacznijmy od prostej, ale o zaskakująco dobrych własnościach, graficznej metody diagnozującej zbieżność do wartości oczekiwanej, jaką jest jest wykorzystanie *sum skumulowanych* – CUSUM (z j. ang. *cumulative sums*). W metodzie tej badany jest wykres wartości

$$C_n(i) = \sum_{k=0}^{i} (h(X_k) - S_n) \tag{6.36}$$

względem zmiennej $i = 1, \ldots, n$, gdzie

$$S_n = \frac{1}{n+1} \sum_{k=0}^{n} h(X_k) \ . \tag{6.37}$$

Autorzy tej metody wskazują na zależność pomiędzy wyglądem wykresu a zachowaniem się ŁM – jeśli algorytm MCMC stosunkowo szybko przemieszcza się po całej przestrzeni stanów $\mathcal{X}$, wykres funkcji $C_n(.)$ oscyluje wokół zera. W przeciwnym przypadku na wykresie $C_n(.)$ dostrzec można gładkie trajektorie, z długimi czasami powrotu do sąsiedztwa zerowej wartości. Niestety, z racji bazowania analizy tylko na pojedynczym łańcuchu, wnioski uzyskane z obserwacji wartości (6.36) mogą być mylące w niektórych przykładach, wskazując błędnie na osiągnięcie zbieżności nawet przy jej braku. Ponadto jest to metoda w swej naturze heurystyczna, wiele więc zależy od intuicji obserwatora. Jej rozszerzenia, poprzez porównywanie do ŁM o ustalonych wcześniej własnościach, są mniej subiektywne.

Inną metodą jest jednoczesny monitoring kilku różnych estymatorów średniej $\mathbb{E}_f h(X)$ opartych na tej samej trajektorii. W momencie, w którym wartości owych estymatorów będą sobie dostatecznie bliskie, możemy zdiagnozować zbieżność metody MCMC. Oprócz standardowego estymatora (5.4), w metodzie tej może być wykorzystana specyficzna wersja warunkowego estymatora średniej, czyli estymator Rao-Blackwellizowany (6.15). Estymator taki może zostać również stworzony dla próbnika Gibbsa (patrz np. [30]). Jego ogólną postać zapisać możemy jako

$$\mathbb{E}_f^{\mathrm{RB}} h(X) = \frac{1}{n+1} \sum_{i=0}^{n} \mathbb{E}_f (h(X_i)|Z_i) \ , \tag{6.38}$$

gdzie Z_i są pewnymi dodatkowymi zmiennymi losowymi, np. dla algorytmu MH były to prawdopodobieństwa postaci $\mathrm{P}(X_i = Y \,|Y_0, Y_1, \ldots, Y_n)$.

Kolejnym przybliżeniem wartości średniej może być estymator uzyskany przy pomocy metody próbkowania ważonego (*importance sampling*, porównaj z (5.22)). Jak przedstawiono to wcześniej przy metodach MC, estymator taki obliczamy poprzez generowanie próbek z innej, niż interesująca nas funkcja

gęstości $f(x)$. W przypadku, gdy gęstość $f(x)$ znamy z dokładnością do stałej normującej, możemy stworzyć estymator postaci

$$\mathbb{E}_f^{\mathrm{IS}} h(X) = \frac{1}{n+1} \sum_{i=0}^{n} w_i h(X_i) \ , \tag{6.39}$$

gdzie

$$w_i \propto \frac{f(X_i)}{g_i(X_i)} \tag{6.40}$$

i $g_i(X_i)$ jest prawdziwą gęstością użytą do generowania wartości w kroku X_i.

Estymator $\mathbb{E}_f h(X)$ może wreszcie być stworzony poprzez zastosowanie aproksymacji Riemanna (w j. ang. *Riemann approximation*. Wyraża się on wtedy wzorem

$$\mathbb{E}_f^{\mathrm{R}} h(X) = \sum_{i=0}^{n} \left(X_{(i+1)} - X_{(i)} \right) h(X_{(i)}) f(X_{(i)}) \ , \tag{6.41}$$

gdzie $\left(X_{(i)}\right)_i$ (dla $i = 0, \ldots, n$) oznacza uporządkowany niemalejąco ciąg wyjściowy $X_0, X_1, \ldots, X_n$. Należy zauważyć, że estymator tej postaci jest przeznaczony tylko dla jednowymiarowych zmiennych losowych, co w znacznym stopniu ogranicza jego stosowalność.

Jeśli cztery powyższe estymatory (5.4), (6.38) – (6.41) są w rozważanym przykładzie możliwe do uzyskania, wykorzystać je można jako różne indykatory zbieżności do wartości oczekiwanej. W takim przypadku symulacje kończy się, jeśli wszystkie estymatory uzyskają zgodność do ustalonego miejsca po przecinku.

Choć metoda ta ma silne podstawy teoretyczne, niemniej posiada też kilka istotnych wad. Mianowicie nie we wszystkich przykładach udaje się znaleźć wszystkie ze wspomnianych powyżej estymatorów. Ponadto metoda ta jest bardzo konserwatywna i może prowadzić do znacznego zwiększenia zbędnych obliczeń. Wreszcie, ponieważ wszystkie estymatory bazują tylko na jednym łańcuchu, może zdarzyć się wyciągnięcie na ich podstawie nieprawidłowych wniosków.

Przykładem innej metody diagnostycznej jest jednoczesne badanie estymatorów wariancji pomiędzy wyrazami pojedynczej trajektorii i jednocześnie pomiędzy poszczególnymi trajektoriami. Niech $K > 1$ oznacza ilość łańcuchów, a $X_i^{(k)}$ – i-ty element k-tej trajektorii dla $i = 0, \ldots, n$ oraz $k = 1, \ldots, K$. W takim przypadku estymator dla *wariancji zewnętrznej*, tzn. *pomiędzy łańcuchami*, można przedstawić w postaci

$$\overline{\mathrm{Var}}_{\mathrm{B}} = \frac{1}{K} \sum_{k=1}^{K} \left(\overline{h(X^{(k)})} - \overline{h(X)} \right)^2 \ , \tag{6.42}$$

gdzie

$$\overline{h(X^{(k)})} = \frac{1}{n+1} \sum_{i=0}^{n} h\left(X_i^{(k)}\right) \ , \ \overline{h(X)} = \frac{1}{K} \sum_{k=1}^{K} \overline{h\left(X^{(k)}\right)} \ . \tag{6.43}$$

Z kolei estymator *wariancji wewnętrznej*, tzn. *wewnątrz pojedynczych łańcuchów*, zapisać można jako

$$\overline{\mathrm{Var}}_{\mathrm{W}} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{n+1} \sum_{i=0}^{n} \left(h\left(X_i^{(k)}\right) - \overline{h\left(X^k\right)} \right)^2 \right) \ . \tag{6.44}$$

Rozważać także można estymator wariancji postaci

$$\overline{\mathrm{Var}}' = \frac{n}{n+1}\overline{\mathrm{Var}}_{\mathrm{W}} + \overline{\mathrm{Var}}_{\mathrm{B}} \tag{6.45}$$

i w celu diagnozowania zbieżności porównywać go z estymatorem $\overline{\mathrm{Var}}_{\mathrm{W}}$. Oba te estymatory są bowiem asymptotycznie zbieżne, tak wiec zmniejszenie się różnicy pomiędzy nimi poniżej ustalonej wartości może wskazywać na zbieżność algorytmu.

Inną metodą diagnozy zbieżności jest monitorowanie statystyki

$$R = \frac{\overline{\mathrm{Var}}' + \frac{\overline{\mathrm{Var}}_{\mathrm{B}}}{K}}{\overline{\mathrm{Var}}_{\mathrm{W}}} \, \frac{\nu}{\nu - 2} \, , \tag{6.46}$$

gdzie

$$\nu = 2\frac{\left(\overline{\mathrm{Var}}'^2 + \frac{\overline{\mathrm{Var}}_{\mathrm{B}}}{K}\right)^2}{\overline{\mathrm{Var}}_{\mathrm{W}}} \tag{6.47}$$

i stworzeniu na jej podstawie przedziału ufności opartym na przybliżeniu rozkładem Fishera lub sprawdzaniu zbieżności do jedynki.

6.8.3 Inne kryteria i metody diagnozy zbieżności

Oprócz pokrótce przedstawionych powyżej kryteriów i metod zbieżności, znane są również inne, użyteczne wyniki na tym polu.

Jednym z takich innego typu kryteriów jest np. badanie, w jakim stopniu obserwacje $X_0, X_1, \ldots$ możemy traktować jako próbę *iid*. Podejście takie umożliwia łatwe ominięcie problemów związanych z estymacją wariancji w twierdzeniach ergodycznych dla ŁM (patrz dyskusja w rozdziale 6.6).

Przykładami innych metod diagnostycznych są wykorzystanie metody *sprzęgania* (lub *parowania*, w j. ang. *coupling*), estymowanie normy *wahania całkowitego* (lub *zmienności całkowitej*, w j. ang. *total variation*) dla rozkładu otrzymanego z symulacji, wykorzystanie pewnych własności prawdopodobieństw przejścia pomiędzy krokami w ŁM, czy estymatorów uzyskanych heurystycznie z symulacji pomocniczych.

Należy zauważyć, że metody diagnozy wykorzystywane w metodach MCMC różnią się między sobą nie tylko pod względem kryteriów zbieżności. Bardzo istotnym czynnikiem jest też wykorzystywanie przez niektóre metody obserwacji *jednocześnie z wielu trajektorii*, podczas gdy inne zadowalają się informacjami uzyskanymi tylko z *jednego łańcucha*. Przykładem metody pierwszego typu jest badanie estymatorów wariancji pomiędzy i wewnątrz łańcuchów, opisane w rozdziale 6.8.2.

W literaturze dotyczącej metod MCMC istnieje długotrwała dyskusja dotycząca wad i zalet obu wymienionych powyżej sposobów, tzn. konstruowania kilku krótszych trajektorii lub tylko jednego, ale za to dłuższego łańcucha. W szczególności korzystanie jedynie z pojedynczego łańcucha może prowadzić do obciążenia estymatorów poszukiwanych metodą MCMC.

Rozdział 7

Resampling

W poniższym rozdziale pokrótce omówimy niektóre metody *resamplingu*, bardziej szczegółowo opisane w [9, 10]. Ogólnie rzecz biorąc, algorytmy to polegają na ponownym wykorzystaniu pobranej próby (w sensie próby statystycznej) $X_1, X_2, \ldots, X_n$.

7.1 Bootstrap

Bootstrap zaproponowany został w [8]. Ideę tej metody przedstawimy na klasycznym przykładzie zaczerpniętym z [10].

Przykład 7.1. *Przypuśćmy, że grupę 16 myszy laboratoryjnych podzielono losowo na dwie grupy. Pierwszą z nich, złożoną z 7 myszy, poddano działaniu nowego leku. Drugą grupę myszy (pozostałych 9) przydzielono do tzw. grupy kontrolnej, której nie poddano działaniu nowego leku. Dla grupy poddanej działaniu leku otrzymano następujące dane dotyczące czasu przeżycia*

$$94 \;\; 197 \;\; 16 \;\; 38 \;\; 99 \;\; 141 \;\; 23 \; .$$

W grupie kontrolnej odpowiednie dane miały postać

$$52 \;\; 104 \;\; 146 \;\; 10 \;\; 51 \;\; 30 \;\; 40 \;\; 27 \;\; 46 \; .$$

Czy podany lek zwiększył czas przeżycia myszy?

Rozwiązanie: Przy tak postawionym pytaniu niezbędne jest wykonanie odpowiedniego testu statystycznego. Przyjrzyjmy się najpierw średnim czasom przeżycia dla pierwszej grupy

$$\bar{x} = \frac{94 + \ldots + 23}{7} = 86,86 \tag{7.1}$$

i dla drugiej grupy

$$\bar{y} = \frac{52 + \ldots + 46}{9} = 56,22 \; . \tag{7.2}$$

Różnica pomiędzy średnimi wynosi 30,63, co zdaje się sugerować nierówność średnich i istotne działanie lekarstwa. Problemem jest jednak *zróżnicowanie* danych w każdej z grup, które wpływa na *losowy błąd pomiaru*, a – co za tym

idzie – na interpretację różnicy średnich. Estymator błędu standardowego jest w przypadku średniej dany wzorem

$$\hat{se}_{\bar{x}} = \sqrt{\frac{s_0^2}{n}} \ , \tag{7.3}$$

gdzie s_0^2 jest wariancją próbkową, tzn.

$$s_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ . \tag{7.4}$$

W ten sposób dla naszych danych

$$\hat{se}_{\bar{x}} = 25,24 \ , \ \hat{se}_{\bar{y}} = 14,14 \ , \ \hat{se}_{\bar{x}-\bar{y}} = \sqrt{25,24^2 + 14,14^2} = 28,93 \ . \tag{7.5}$$

Wynik ten wskazuje, że błąd oszacowania różnicy średnich jest duży w porównaniu do samej różnicy średnich. To powoduje, że przy pewnych poziomach istotności, przyjęta zostanie hipoteza o równości średnich, czyli braku działania leku. ◇

Zamiast średnich i ich różnicy, można spróbować zastosować inne statystyki próbki w celu wykazania działania leku lub przyjęcia hipotezy o braku jego działania. Przykładem może być tutaj obliczenie mediany dla pierwszej i drugiej grupy obserwacji. Niestety, dla dowolnej ogólnej statystyki $S(\mathbf{x})$ bazującej na próbce $\mathbf{x} = x_1, x_2, \ldots, x_n$ może nie istnieć prosty wzór na oszacowanie błędu standardowego, jaki podaliśmy dla (7.3). W takim przypadku pomocny może być boostrap.

Załóżmy, że dysponujemy próbą $\mathbf{x} = x_1, x_2, \ldots, x_n$, pochodzącą z obserwacji. W metodzie bootstrapu generujemy *próby bootstrapowe* (*bootstrap sample*) $\mathbf{X}^{*1}, \mathbf{X}^{*2}, \ldots, \mathbf{X}^{*B}$. Próba $\mathbf{X}^{*j} = X_1^{*j}, X_2^{*j}, \ldots, X_n^{*j}$ powstaje poprzez losowanie ze zwracaniem n elementów z oryginalnej, zaobserwowanej wcześniej próby $\mathbf{x}$. Prawdopodobieństwo wylosowania każdego z elementów z próby $\mathbf{x}$ jest przy tym takie samo i wynosi $\frac{1}{n}$.

W następnym kroku obliczana jest wartość szukanej statystyki $S(.)$ dla każdej z realizacji prób bootstrapowych, otrzymujemy zatem $S(\mathbf{x}^{*1}), \ldots, S(\mathbf{x}^{*B})$. Boostrapowy estymator błędu standardowego statystyki $S(.)$ jest w takim przypadku dany próbkowym odchyleniem standardowym dla prób standardowych, czyli

$$\sqrt{\frac{\sum_{b=1}^{B} \left(S(\mathbf{x}^{*b}) - \bar{S}(\mathbf{x}^{*})\right)^2}{B-1}} \ , \tag{7.6}$$

gdzie

$$\bar{S}(\mathbf{x}^{*}) = \frac{\sum_{b=1}^{B} S(\mathbf{x}^{*b})}{B} \ . \tag{7.7}$$

Jak widzimy, bootstrap służy tutaj do prostego, symulacyjnego oszacowania błędu standardowego statystyki $S(.)$. Jest to przydatne w częstych przypadkach, w których prostego wzoru analitycznego na ów błąd standardowy nie ma.

Odpowiedni algorytm ma zatem postać:

Algorytm 7.2.

1. Dla próby $\mathbf{x}$ zbuduj dystrybuantę empiryczną $\hat{F}$

2. Wygeneruj próbę bootstrapową $\mathbf{X}^{*j}$ z dystrybuanty $\hat{F}$, tzn.

$$X_1^{*j}, X_2^{*j}, \ldots, X_n^{*j} \overset{iid}{\sim} \hat{F} \tag{7.8}$$

   Krok ten wykonaj B razy, niezależnie dla $j = 1, 2, \ldots, B$.

3. Dla każdej z realizacji próby bootstrapowej $\mathbf{x}^{*j}$ oblicz wartość szukanej statystyki $S(\mathbf{x}^{*j})$

4. Odchylenie standardowe $S(\mathbf{x}^{*1}), S(\mathbf{x}^{*2}), \ldots, S(\mathbf{x}^{*B})$ dane jest wzorem

$$s_{*,S(.)} = \sqrt{\frac{\sum_{b=1}^{B} \left(S(\mathbf{x}^{*b}) - \bar{S}(\mathbf{x}^{*})\right)^2}{B-1}} \ , \tag{7.9}$$

   gdzie

$$\bar{S}(\mathbf{x}^{*}) = \frac{\sum_{b=1}^{B} S(\mathbf{x}^{*b})}{B} \ . \tag{7.10}$$

Przy $B \to \infty$ odchylenie standardowe próbkowe $s_{*,S(.)}$ zbiega do $s^*_{S(.)}(\hat{F})$, czyli bootstrapowego estymatora odchylenia standardowego statystyki $S(.)$. W celu ominięcia konieczności przejścia granicznego, zazwyczaj (patrz [9, 10]) generuje się 100 – 200 prób bootstrapowych, co wystarcza do osiągnięcia zadowalającej dokładności.

Oczywiście bootstrap zastosować można do bardziej ogólnych problemów, niż tylko estymacja odchylenia standardowego, czy błędu standardowego. W takim przypadku w kroku 4 zamiast odchylenia standardowego $s_{*,S(.)}$ znajdujemy poszukiwaną funkcję statystyki $S(.)$, np. jej wartość oczekiwaną lub prawdopodobieństwo pewnego zdarzenia.

7.2 Jackknife

Metoda *jackknife* zaproponowana przez Quenouille'a (patrz [29]) również zakłada ponowne wykorzystanie otrzymanej realizacji próby $\mathbf{x}$. Zamiast jednak losowania z dystrybuanty empirycznej $\hat{F}$, wykorzystuje się ciąg realizacji z kolejno wykasowanymi, pojedynczymi wartościami.

Załóżmy, że interesuje nas pewna funkcja $g(.)$, której wartości zależą od dystrybuanty rozkładu $F(.)$, czyli funkcja postaci $g(F)$. Przykładem $g(F)$ może być wartość oczekiwana, kwantyle, itd. Bazując na standardowej próbie $\mathbf{X} = X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, nieznaną wartość funkcji $g(F)$ estymujemy za pomocą statystyki

$$\hat{g} = g(\hat{F}) \ , \tag{7.11}$$

czyli funkcji zależnej od dystrybuanty empirycznej $\hat{F}(\mathbf{x})$. Postać (7.11) zapewnia, że $\hat{g}$ jest niezmiennicza względem permutacji realizacji $\mathbf{x} = x_1, x_2, \ldots, x_n$. Obciążeniem estymatora $\hat{g}$ nazywać tutaj będziemy

$$\mathbf{B} = \mathbb{E}_F \, g(\hat{F}) - g(F) \ . \tag{7.12}$$

Metoda jackknife polega na wygenerowaniu ciągu realizacji, w których wykasowano pojedynczą, kolejną wartość, tzn.

$$\begin{aligned}\mathbf{x}^{(-1)} &= (x_2, x_3, \ldots, x_n), \mathbf{x}^{(-2)} = (x_1, x_3, \ldots, x_n), \ldots, \\ \mathbf{x}^{(-j)} &= (x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n), \ldots, \mathbf{x}^{(-n)} = (x_1, x_2, \ldots, x_{n-1}) \ ,\end{aligned} \tag{7.13}$$

któremu odpowiada ciąg dystrybuant empirycznych

$$\hat{F}^{(-j)} = x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n \tag{7.14}$$

dla $j = 1, 2, \ldots, n$. Na podstawie $\mathbf{x}^{(-1)}, \ldots, \mathbf{x}^{(-n)}$ obliczamy odpowiednie estymatory funkcji $g(.)$, tzn.

$$\hat{g}^{(-j)} = g\left(\hat{F}^{(-j)}\right) = g\left(x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n\right) \tag{7.15}$$

dla $j = 1, 2, \ldots, n$.

Estymatorów postaci (7.15) możemy użyć do estymowania obciążenia (7.12). W tym celu Quenouille zaproponował estymator obciążenia $\mathbf{B}$ dany wzorem

$$\hat{\mathbf{B}} = (n-1)\left(\hat{g}^{(.)} - \hat{g}\right) \ , \tag{7.16}$$

gdzie

$$\hat{g}^{(.)} = \frac{1}{n}\sum_{i=1}^{n} \hat{g}^{(-i)} \ . \tag{7.17}$$

Prowadzi to do *estymatora metody jackknife* dla funkcji g, który jest poprawiony o estymator obciążenia (7.16)

$$\hat{g}_{\text{jack}} = \hat{g} - \hat{\mathbf{B}} = n\hat{g} - (n-1)\hat{g}^{(.)} \ . \tag{7.18}$$

Estymator (7.16) jest poprawnym przybliżeniem obciążenia np. dla wariancji

$$\hat{g} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \ , \tag{7.19}$$

która jest przypadkiem funkcjonału kwadratowego, tzn. $\hat{g}$ może być przedstawione w postaci

$$\hat{g} = \mu + \frac{1}{n}\sum_{i=1}^{n} \alpha(x_i) + \frac{1}{n^2}\sum_{1 \leqslant i_1 < i_2 \leqslant n} \beta\left(x_{i_1}, x_{i_2}\right) \ , \tag{7.20}$$

czyli funkcji bez wyrazów wyższych rzędów zależnych od x_i. Mówi o tym następujące twierdzenie:

Twierdzenie 7.3. *Dla funkcjonałów kwadratowych, estymator* (7.16) *jest nieobciążonym estymatorem obciążenia* (7.12).

Głównym celem estymatora metody jackknife (7.18) jest w takim przypadku zmniejszenie błędu otrzymywanego przybliżenia z $O(1/n)$ na $O(1/n^2)$.

Metoda jackknife może zostać wykorzystana również w innych przypadkach, np. do nieparametrycznej estymacji wariancji estymatora $\hat{g}$, czyli

$$\mathbf{V} = \mathbb{E}_F \left(\hat{g}(\mathbf{X}) - \mathbb{E}_F\, \hat{g}\right)^2 \ . \tag{7.21}$$

W [35] zaproponowano w tym celu estymator postaci

$$\hat{\mathbf{V}} = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{g}^{(-i)} - \hat{g}^{(\cdot)}\right)^2 \ . \tag{7.22}$$

7.3 Uogólnienie podejście

Na omówione w poprzednich rozdziałach procedury resamplingu można spojrzeć w szerszy sposób, dostrzegając ich podobieństwa i możliwe uogólnienia. Niech p^* będzie wektorem o n możliwych wartościach, t.ż.

$$p^*(X_i) \geqslant 0, \text{ dla } i = 1, 2 \ldots, n \ , \ \sum_{i=1}^{n} p^*(X_i) = 1 \ . \tag{7.23}$$

Wtedy dla próby $\mathbf{X} = X_1, X_2, \ldots, X_n$ i dla dowolnego wektora p^* otrzymujemy *ważoną empiryczną dystrybuantę* $\hat{F}^*$ daną wzorem

$$\hat{F}^*(x) = \sum_{i=1}^{n} \mathbb{1}(X_i \leqslant x) p^*(X_i) \ . \tag{7.24}$$

Jak widzimy, ważona dystrybuanta empiryczna różni się od zwykłej dystrybuanty empirycznej (1.38) tym, że z każdą wartością X_i nie jest związana masa atomowa $\frac{1}{n}$, ale pewne prawdopodobieństwo $p^*(X_i)$, być może równe zero.

Wektor p^* i dystrybuanta $\hat{F}^*$ definiują odpowiednie procedury resamplingu realizacji próby $\mathbf{x} = x_1, x_2, \ldots, x_n$. W przypadku bootstrapu wektor ten ma postać

$$p^*(X_1) = \frac{1}{n}, p^*(X_2) = \frac{1}{n}, \ldots, p^*(X_n) = \frac{1}{n}. \tag{7.25}$$

Jak widzieliśmy, dla wektora (7.25) tworzona jest dystrybuanta empiryczna, z której następuje losowanie prób bootstrapowych. Natomiast dla metody jackknife jest to ciąg wektorów postaci

$$\begin{aligned} p^*_{(j)}(X_1) = \frac{1}{n-1}, p^*_{(j)}(X_2) = \frac{1}{n-1}, \ldots, \\ p^*_{(j)}(X_{j-1}) = \frac{1}{n-1}, p^*_{(j)}(X_{j+1}) = \frac{1}{n-1}, \ldots, p^*_{(j)}(X_n) = \frac{1}{n-1} \end{aligned} \tag{7.26}$$

dla którego obliczamy odpowiednie estymatory $\hat{g}^{(-j)}$, przy $j = 1, 2, \ldots, n$.

Bibliografia

[1] Besag J. E., *On the statistical analysis of dirty pictures (with discussion)*, J. R. Statist. Soc. B, 48, pp. 259 – 302, 1986

[2] Binder K., *Monte Carlo Simulations in Statistical Physics: An Introduction*, Springer, 1988

[3] Birkohlc A., *Analiza matematyczna – funkcje wielu zmiennych*, PWN, Warszawa 1986

[4] Bremaud P., *Markov Chains — Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer–Verlag, New York, 1999

[5] Cappé O., Robert Ch. P., *Markov Chain Monte Carlo: 10 Years and Still Running!*, Journal of the American Statistical Association, December 2000, Vol. 95, No. 452

[6] Casella G., Robert C. P., *Rao-Balckwellisation of sampling schemes*, Biometrika, Vol. 83, No. 1, pp. 81 – 94, 1996

[7] Eckhardt R., *Stan Ulam, John von Neumann, and the Monte Carlo Method*, Los Alamos Science, Special Issue (15), 131 – 137, 1987

[8] Efron B., *Bootstrap methods: another look at the jackknife*, Ann. Statist., 7, 1 – 26, 1979

[9] Efron B., *The jackknife, the bootstrap, and other resampling plans*, Philadelphia: Pa. Society for Industrial and Applied Mathematics, 1982

[10] Efron B., Tibshirani R. J., *An Introduction to the Bootstrap*, Chapman & Hall, 1993

[11] Feller W., *An Introduction to Probability Theory and its Applications*, Vol. 1, John Wiley, New York, 1970

[12] Fortuna Z., Macukow B., Wąsowski J., *Metody numeryczne*, WNT 1998

[13] Geman S., MacClure D. E., *Statistical methods for tomographic image reconstruction*, Bull. Int. Statist. Inst., LII-4, pp. 5 – 21, 1987

[14] Geyer C. J., *Practical Markov chain Monte Carlo (with discussion)*, Statist. Sci. 7, 473–511, 1992

[15] Gilks W. R., Richardson S., Spiegelhalter D. J., *Markov Chain Monte Carlo in Parctice*, Chapman & Hall, 1997

[16] Glasserman P., *Monte Carlo Methods in Financial Enigineering*, Springer-Verlag, New York 2004

[17] Hryniewicz O., *Wykłady ze statystyki dla studentów informatycznych technik zarządzania*

[18] Hull J. C., *Options, Futures and Other Derivatives*, Presntice Hall, 1997

[19] Jakubowski J., Sztencel R., *Wstęp do teorii prawdopodobieństwa*, Script, Warszawa 2000

[20] Kipnis C., Varadhan S. R., *Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions*, Comm. Math. Phys., 104, 1–19, 1986

[21] Koronacki J., Lasota S., Niemiro W., *Positron emission tomography by Markov chain Monte Carlo with auxiliary variables*, Pattern Recognition, 38, 241 – 250, 2005

[22] Lasota S., Niemiro W., *A version of the Swendsen-Wand algorithm for restoration of images degraded by Poisson noise*, Pattern Recognition, 36, 931 – 941, 2003

[23] Lauritzen S. L., Dawid A. P., Larsen B. N., Leimer H.-G., *Independence properties of directed Markov fields*, Networks, 20, pp. 491 – 505, 1990

[24] Liu J., Wong W., Kong A., *Correlaction structure and convergence rate of the Gibbs sampler with various scans*, J. Royal Statist. Soc. Series B, 57, 157 – 169, 1995

[25] Metropolis N., Rosenbluth A.W., Rosenbluth M. N., Teller A. H., Teller E., *Equations of state calculations by fast computing machines*, J. Chem. Phys. 21, 1953

[26] Metropolis N., Ulam S., *The Monte Carlo Method*, Journal of American Statistical Association, 44, 1949

[27] Meyn S. P., Tweedie R. L., *Markov Chains and Stochastic Stability*, Springer Verlag, New York, 1993

[28] Nummelin E., *MC's for MCMC'ists*, Preprint 310, December 2001

[29] Quenouille M., *Approximate tests of correlation in time series*, J. Roy. Statist. Soc. Ser. B, 11, 18 – 84, 1949

[30] Robert Ch. P., Casella G., *Monte Carlo Statistical Methods*, Springer–Verlag, 2nd ed., New York, 2004

[31] Roberts G. O., Rosenthal J. S., *Small and pseudo - small sets for Markov Chains*, Stochastic Models, Vol. 17, No. 2, 2001

[32] Rubinstein R. Y., *Simulation and the Monte Carlo Method*, J. Wiley, New York, 1981

[33] Thomas A., Spiegelhalter D. J., Gilks W. R., *BUGS: a Program to Perform Bayesian Inference using Gibbs Sampling* w Bernardo J. M., Berger J. O., Dawid A. P., Smith A. F. M. (eds.) *Bayesian Statistics 4*, Oxford University Press 1992

[34] Tierney L., *Markov chains for exploring posterior distributions (with discussion)*, Ann. Statist., 22, 1701 – 1786, 1994

[35] Tukey J., *Bias and confidence in not quite large samples*, Ann. Math. Statist., 29, 614, 1958

[36] Weron A., Weron R. *Inżynieria finansowa*, WNT 1999, Warszawa

[37] Wieczorkowski R., Zieliński R., *Komputerowe generatory liczb losowych*, WNT, 1997

[38] Yakowitz S., Krimmel J., Szidarovszky F., *Weighted Monte Carlo integration*, SIAM J. Numer. Anal., 15(6), 1289 – 1300, 1978