Matematyka stosowana

Symulacje stochastyczne i metody Monte Carlo

Wojciech NIEMIRO
wniem@mimuw.edu.pl
http://adres.strony1.www

Uniwersytet Warszawski, 2013

Streszczenie. Jest to wykład na temat symulacji zjawisk losowych. Obejmuje też wstęp do metod Monte Carlo (MC), czyli algorytmów zrandomizowanych. Symulacje komputerowe są nie tylko prostym i skutecznym narzędziem badania procesów losowych, ale też znajdują zastosowania do obliczeń wielkości deterministycznych. Metody Monte Carlo a w szczególności algorytmy MCMC, oparte na łańcuchach Markowa należą do standardowych narzędzi obliczeniowych, między innymi w statystyce Bayesowskiej i fizyce statystycznej. Przedmiot jest przeznaczony dla wszystkich studentów lubiących rachunek prawdopodobieństwa, matematyków i informatyków. Powinien zachęcić słuchaczy do „zobaczenia losowości" – przy wykorzystaniu pięknego i dostępnego za darmo pakietu R. Wykład jest utrzymany na elementarnym poziomie.
W pierwszej części wykładu omówione zostaną sposoby generowania zmiennych losowych o zadanym rozkładzie prawdopodobieństwa i prostych procesów stochastycznych. Druga część poświęcona będzie ogólnym zasadom konstrukcji algorytmów Monte Carlo, szacowania ich dokładności i redukcji błędu. Sporo miejsca zajmują Markowowskie algorytmy Monte Carlo, MCMC.

Wersja internetowa wykładu:
http://mst.mimuw.edu.pl/lecture.php?lecture=sst

(może zawierać dodatkowe materiały)

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Projekt współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego.

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Skład w systemie LaTeX, z wykorzystaniem m.in. pakietów beamer oraz listings. Szablony podręcznika i prezentacji: Piotr Krzyżanowski; koncept: Robert Dąbrowski.

Spis treści

1. Wprowadzenie

1.1. Od autora

Jest kilka ważnych powodów, dla których warto się zająć symulacjami stochastycznymi:

— Symulacje stochastyczne są prostym sposobem badania zjawisk losowych.
— Ściśle związane z symulacjami stochastycznymi są metody obliczeniowe nazywane „Monte Carlo" (MC). Polegają one na wykorzystaniu „sztucznie generowanej" losowości w celu rozwiązania zadań deterministycznych. Metody MC są proste i skuteczne. Dla pewnych problemów MC jest jedynym dostępnym narzędziem obliczeniowym. Dla innych problemów MC jest co prawda mniej efektywne od metod numerycznych, ale za to dużo łatwiejsze!
— W moim przekonaniu symulacje stochastyczne są wspaniałą pomocą przy nauce rachunku prawdopodobieństwa. Pozwalają lepiej „zrozumieć losowość".
— Symulacje stochastyczne są dostępne dla każdego. W szczególności, „otoczenie" **R**, które stanowi naprawdę potężne narzędzie, jest rozpowszechniane **za darmo!**

Jest wreszcie powód najważniejszy:

— Symulacje stochastyczne są **świetną zabawą!**

Literatura na temat symulacji stochastycznych jest bardzo obszerna. Godna polecenia jest książka Zielińskiego i Wieczorkowskiego [23], poświęcona w całości generatorom zmiennych losowych. Przedstawia ona bardziej szczegółowo zagadnienia, odpowiadające Rozdziałom 2–4 niniejszego skryptu i zawiera materiał, który zdecydowałem się pominąć: wytwarzanie „liczb losowych" o rozkładzie jednostajnym i testowanie generatorów. Podobne zagadnienia są przedstawione trochę w innym stylu w monografii Ripleya [18], która również zawiera wstęp do metod Monte Carlo. Zaawansowane wykłady można znaleźć w nowoczesnych monografiach Asmussena i Glynna [2], Liu [15], Roberta i Caselli [19]. Pierwsza z nich jest zorientowana bardziej na wyniki teoretyczne, zaś druga bardziej na zastosowania. Świetnym wstępem do metod MCMC są prace Geyera [7] i [8]. Teoria łańcuchów Markowa z uwzględnieniem zagadnień istotnych dla MCMC jest przystępnie przedstawiona w książce Brémaud [4]. Podstawy teoretycznej analizy zrandomizowanych algorytmów (tematy poruszane w Rozdziale 15 skryptu) są znakomicie przedstawione w pracach Jerruma i Sinclaira [11] oraz Jerruma [12].

1.2. Przykłady

Zacznę od kilku przykładów zadań obliczeniowych, które można rozwiązywać symulując losowość. Wybrałem przykłady najprostsze, do zrozumienia których wystarcza zdrowy rozsądek i nie potrzeba wielkiej wiedzy.

Przykład 1.1 (Igła Buffona, 1777)**.** Podłoga jest nieskończoną płaszczyzną, podzieloną równoległymi prostymi na „deski" szerokości d. Rzucamy „losowo" igłę o długości l. Elementarne rozważania prowadzą do wniosku, że

$$p = \mathbb{P}(\text{igła przetnie którąś z prostych}) = \frac{2l}{\pi d}.$$

Buffon zauważył, że tę prostą obserwację można wykorzystać do... obliczania liczby π metodą „statystyczną". Powtórzmy nasze doświadczenie niezależnie n razy. Oczywiście, mamy do czynienia ze schematem Bernoulliego, w którym „sukcesem" jest przecięcie prostej. Niech

$$\begin{aligned}\hat{p}_n &= \frac{\text{liczba sukcesów}}{\text{liczba doświadczeń}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\text{w } i\text{-tym doświadczeniu igła przecięła prostą})\end{aligned}$$

Wielkość $\hat{p}_n$ jest empirycznym odpowiednikiem prawdopodobieństwa p i powinna przybliżać to prawdopodobieństwo, przynajmniej dla dużych n. W takim razie, możemy za przybliżenie liczby π przyjąć

$$\pi_n = \frac{2l}{\hat{p}_n d}.$$

Statystyk powiedziałby, że *zmienna losowa* π_n jest *estymatorem* liczby π. To wszystko jest bardzo proste, ale parę kwestii wymaga uściślenia. Jak duża ma być liczba powtórzeń, żeby przybliżenie było odpowiednio dokładne? Ale przecież mamy do czynienia ze „ślepym losem"! Czyż pech nie może sprawić, że mimo dużej liczby doświadczeń przybliżenie jest kiepskie? Czy odpowiednio dobierając l i d możemy poprawić dokładność? A może da się zaprojektować lepsze doświadczenie?

Przykład 1.2 (Sieć zawodnych połączeń)**.** Niech $\mathcal{V}, \mathcal{E}$ będzie grafem skierowanym spójnym. Krawędzie reprezentują „połączenia" pomiędzy wierzchołkami. Krawędź $e \in \mathcal{E}$, niezależnie od pozostałych, ulega awarii ze znanym prawdopodobieństwem p_e. W rezultacie powstaje losowy podzbiór $C \subseteq \mathcal{E}$ sprawnych połączeń o rozkładzie prawdopodobieństwa

$$P(C) = \prod_{e \notin C} p_e \prod_{e \in C} (1 - p_e).$$

Jest jasne, jak można symulować to zjawisko: dla każdej krawędzi e „losujemy" zmienną U_e o rozkładzie równomiernym na $[0, 1]$ i przyjmujemy

$$\begin{cases} e \notin C & \text{jeśli } U_e \leqslant p_e; \\ e \in C & \text{jeśli } U_e > p_e. \end{cases}$$

Powiedzmy, że interesuje nas możliwość znalezienia ścieżki (ciągu krawędzi) wiodącej z ustalonego wierzchołka v_0 do innego wierzchołka v_1. Niech

$$\theta = \mathbb{P}(\text{ w zbiorze } C \text{ istnieje ścieżka z } v_0 \text{ do } v_1).$$

Generujemy niezależnie n kopii $C_1, \ldots, C_n$ zbioru C. Nieznane prawdopodobieństwo θ przybliżamy przez odpowiednik próbkowy:

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\text{ w zbiorze } C_i \text{ istnieje ścieżka z } v_0 \text{ do } v_1).$$

Przykład 1.3 (Skomplikowana całka)**.** Załóżmy, że $X_1, \ldots, X_m$ są niezależnymi zmiennymi losowymi o jednakowym rozkładzie $\mathrm{N}(0,1)$. Niech

$$R = \max_{k=1}^{m} \sum_{i=1}^{k} X_i - \min_{k=1}^{m} \sum_{i=1}^{k} X_i.$$

Chcemy obliczyć dystrybuantę zmiennej losowej R,

$$H(x) = \mathbb{P}(R \leqslant x).$$

Zauważmy, że z definicji,

$$H(x) = \int \cdots \int_{\Omega_m} (2\pi)^{-m/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{m} x_i^2\right] \mathrm{d}x_1 \cdots \mathrm{d}x_m,$$

gdzie $\Omega_m = \left\{(x_1, \ldots, x_m) : \max_{k=1}^{m} \sum_{i=1}^{k} x_i - \min_{k=1}^{m} \sum_{i=1}^{k} x_i\right\}$. W zasadzie, jest to więc zadanie obliczenia całki. Jednak skomplikowany kształt wielowymiarowego zbioru Ω_m powoduje, że zastosowanie standardowych metod numerycznych jest utrudnione. Z kolei symulowanie zmiennej losowej R jest bardzo łatwe, wprost z definicji. Można wygenerować wiele niezależnych zmiennych $R_1, \ldots, R_n$ o rozkładzie takim samym jak R. Wystarczy teraz policzyć ile spośród tych zmiennych jest $\leqslant x$:

$$\hat{H}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(R_i \leqslant x).$$

Schemat symulacyjnego obliczania prawdopodobieństwa jest taki sam jak w dwu poprzednich przykładach. Podobnie zresztą jak w tamtych przykładach, podstawowy algorytm można znacznie ulepszać, ale dyskusję na ten temat odłóżmy na później.

Przykład 1.4 (Funkcja harmoniczna)**.** Następujące zadanie jest dyskretnym odpowiednikiem sławnego zagadnienia Dirichleta. Niech $\mathcal{D}$ będzie podzbiorem kraty całkowitoliczbowej $\mathbb{Z}^2$. Oznaczmy przez $\partial\mathcal{D}$ brzeg tego zbioru, zdefiniowany następująco:

$$\begin{aligned}(x,y) \in \partial\mathcal{D} \equiv (x,y) \notin \mathcal{D} \text{ i } \big[&(x+1,y) \in \mathcal{D} \text{ lub } (x-1,y) \in \mathcal{D} \\ &\text{lub } (x,y+1) \in \mathcal{D} \text{ lub } (x,y-1) \in \mathcal{D}\big].\end{aligned}$$

Powiemy, że $u : \mathcal{D} \cup \partial\mathcal{D} \to \mathbb{R}$ jest funkcją harmoniczną, jeśli dla każdego punktu $(x,y) \in \mathcal{D}$,

$$u(x,y) = \frac{1}{4}\sum_{(x',y') \in \partial(x,y)} u(x',y'),$$

gdzie sumowanie rozciąga się na 4 punkty (x',y') sąsiadujące z (x,y), to znaczy $\partial(x,y) = \{(x+1,y),(x-1,y),(x,y+1),(x,y-1)\}$.

Mamy daną funkcję na brzegu: $\bar{u} : \partial\mathcal{D} \to \mathbb{R}$. Zadanie polega na skonstruowaniu jej rozszerzenia harmonicznego, to znaczy takiej funkcji harmonicznej $u : \mathcal{D} \cup \partial\mathcal{D} \to \mathbb{R}$, że $u(x,y) = \bar{u}(x,y)$ dla $(x,y) \in \partial\mathcal{D}$.

Wyobraźmy sobie błądzenie losowe po kracie, startujące w punkcie $(x,y) \in \mathcal{D}$. Formalnie jest to ciąg losowych punktów określonych następująco:

$$\begin{aligned}&(X_0, Y_0) = (x,y);\\ &(X_{k+1}, Y_{k+1}) = (X_k, Y_k) + (\xi_{k+1}, \eta_{k+1}),\end{aligned}$$

gdzie (ξ_k, η_k) są niezależnymi wektorami losowymi o jednakowym rozkładzie prawdopodobieństwa:

$$\mathbb{P}\left((\xi_k,\eta_k)=(0,1)\right)=\mathbb{P}\left((\xi_k,\eta_k)=(0,-1)\right)$$
$$=\mathbb{P}\left((\xi_k,\eta_k)=(1,0)\right)=\mathbb{P}\left((\xi_k,\eta_k)=(-1,0)\right)=\frac{1}{4}.$$

Błądzimy tak długo, aż natrafimy na brzeg obszaru. Formalnie, określamy moment zatrzymania $T=\min\{k:(X_k,Y_k)\in\partial\mathcal{D}\}$. Latwo zauważyć, że funkcja

$$u(x,y):=\mathbb{E}[\bar{u}(X_T,Y_T)|(X_0,Y_0)=(x,y)]$$

jest rozwiązaniem zagadnienia! Istotnie, ze wzoru na prawdopodobieństwo całkowite wynika, że

$$u(x,y)=\frac{1}{4}\sum_{(x',y')\in\partial(x,y)}\mathbb{E}[\bar{u}(X_T,Y_T)|(X_1,Y_1)=(x',y')].$$

Wystarczy teraz spostrzeżenie, że rozkład zmiennej losowej $u(X_T,Y_T)$ pod warunkiem $(X_1,Y_1)=(x',y')$ jest taki sam jak pod warunkiem $(X_0,Y_0)=(x',y')$, bo błądzenie „rozpoczyna się na nowo".

Algorytm Monte Carlo obliczania $u(x,y)$ oparty na powyższym spostrzeżeniu został wynaleziony przez von Neumanna i wygląda następująco:

— Powtórz wielokrotnie, powiedzmy n razy, niezależnie doświadczenie:
„błądź startując startując z (x,y) aż do brzegu; oblicz $\bar{u}(X_T,Y_T)$"
— Uśrednij wyniki n doświadczeń.

Dla bardziej formalnego zapisu algorytmu będę się posługiwał pseudo-kodem, który wydaje się zrozumiały bez dodatkowych objaśnień:

Listing.

```
{ 'Gen' oznacza 'Generuj' }
U := 0;
for j = 1 to n
  begin
  (X, Y) := (x, y);
  while (X, Y) ∈ D
    begin
    Gen (ξ, η) ~ U{(0,1), (0,−1), (1,0), (−1,0)}; [ rozkład jednostajny na zbiorze 4−punktowym ]
    (X, Y) := (X, Y) + (ξ, η)
    end
    U := U + ū(X, Y)
  end
U := U/n
```

Przykład 1.5 (Problem „plecakowy")**.** Załóżmy, że $a=(a_1,\ldots,a_m)^\top$ jest wektorem o współrzędnych naturalnych ($a_i\in\{1,2,\ldots\}$) i $b\in\{1,2,\ldots\}$. Rozważamy wektory $x=(x_1,\ldots,x_m)^\top$ o współrzędnych zero-jedynkowych ($x_i\in\{0,1\}$). Interesuje nas *liczba rozwiązań nierówności*

$$x^\top a=\sum_{i=1}^m x_i a_i\leqslant b,$$

a więc liczność $|\mathcal{X}(b)|$ zbioru $\mathcal{X}(b) = \{x \in \{0,1\}^m : x^\top a \leqslant b\}$. Czytelnik domyśla się z pewnością, skąd nazwa problemu. Dokładne obliczenie liczby rozwiązań jest trudne. Można spróbować zastosować prostą metodę Monte Carlo. Jeśli zmienna losowa X ma rozkład jednostajny na przestrzeni $\{0,1\}^m$, to $\mathbb{P}(X \in \mathcal{X}(b)) = |\mathcal{X}(b)|/2^m$. Wystarczy zatem oszacować to prawdopodobieństwo tak jak w trzech pierwszych przykładach w tym rozdziale. Generowanie zmiennej losowej o rozkłdzie jednostajnym na przestrzeni $\{0,1\}^m$ sprowadza się do przeprowadzenia m rzutów monetą i jest dziecinnie proste. Na czym więc polega problem? Otóż szacowane prawdopodobieństwo może być astronomicznie małe. Dla, powiedzmy $a = (1,\ldots,1)^\top$ i $b = m/3$, to prawdopodobieństwo jest $\leqslant \mathrm{e}^{-m/18}$ (proszę się zastanowić jak uzasadnić tę nierówność). Przeprowadzanie ciągu doświadczeń Bernoulliego z takim prawdopodobieństwem sukcesu jest bardzo nieefektywne – na pierwszy sukces oczekiwać będziemy średnio $\geqslant \mathrm{e}^{m/18}$, co dla dużych m jest po prostu katastrofalne.

Metoda, którą naszkicuję należy do rodziny algorytmów MCMC (Monte Carlo opartych na łańcuchach Markowa). Algorytm jest raczej skomplikowany, ale o ile mi wiadomo jest najefektywniejszym ze znanych sposobów rozwiązania zadania. Bez straty ogólności możemy przyjąć, że $a_1 \leqslant \cdots \leqslant a_m$. Niech $b_0 = 0$ oraz $b_j = \min\{b, \sum_{i=1}^{j} a_i\}$. Rozważmy *ciąg* zadań plecakowych ze zmniejszającą się prawą stroną nierówności, równą kolejno $b = b_m, b_{m-1}, \ldots, b_1, b_0 = 0$. Niech więc $\mathcal{X}(b_j) = \{x \in \{0,1\}^m : x^\top a \leqslant b_j\}$. Zachodzi następujący „wzór teleskopowy":

$$|\mathcal{X}(b)| = |\mathcal{X}(b_m)| = \frac{|\mathcal{X}(b_m)|}{|\mathcal{X}(b_{m-1})|} \cdot \frac{|\mathcal{X}(b_{m-1})|}{|\mathcal{X}(b_{m-2})|} \cdots \frac{|\mathcal{X}(b_1)|}{|\mathcal{X}(b_0)|} \cdot |\mathcal{X}(b_0)|.$$

Oczywiście, $|\mathcal{X}(b_0) = 1|$, a więc możemy uznać, że zadanie sprowadza się do obliczenia ilorazów $|\mathcal{X}(b_{j-1})|/|\mathcal{X}(b_j)|$ (pomijamy w tym miejscu subtelności związane z dokładnością obliczeń). Gdybyśmy umieli efektywnie generować zmienne losowe o rozkładzie jednostajnym na przestrzeni $X(b_j)$, to moglibyśmy postępować w dobrze już znany sposób: liczyć „sukcesy" polegające na wpadnięciu w zbiór $X(b_{j-1}) \subseteq X(b_j)$. Rzecz jasna, możemy losować z rozkładu jednostajnego na kostce $\{0,1\}^m$ i eliminować punkty poza zbiorem $X(b_j)$, ale w ten sposób wpadlibyśmy w pułapkę, od której właśnie uciekamy: co zrobić jeśli przez sito eliminacji przechodzi jedno na $\geqslant \mathrm{e}^{m/18}$ losowań?

Opiszemy pewne wyjście, polegające na zastosowaniu błądzenia losowego po zbiorze $X(b_j)$. Dla ustalenia uwagi przyjmijmy $j = m$, czyli $b_j = b$. Losowy ciąg punktów $X(0), X(1), \ldots, X(n), \ldots$ generujemy rekurencyjnie w taki sposób:

Listing.

```
X(0) := 0;
for n = 0 to ∞
  begin
  X := X(n); [ gdzie X = (X_1, ..., X_m)
  Gen I ~ U{1, ..., m}; [ losujemy indeks do zmiany ]
  Y := X; Y_I = 1 − X_I;
  if Y^⊤ a ⩽ b then X(n + 1) := Y
        else X(n + 1) := X
  end
end
```

Zaczynamy błądzenie w punkcie $0 = (0,\ldots,0)^\top$. Jeśli w chwili n jesteśmy w punkcie X, to próbujemy przejść do nowego punktu Y, utworzonego przez zmianę jednej, losowo wybranej współrzędnej punktu X (0 zamieniamy na 1 lub z 1 na 0; pozostałe współrzędne zostawiamy bez zmian). Jeśli „proponowany punkt Y nie wypadł" z rozważanej przestrzeni $\mathcal{X}(b)$, to przechodzimy do punktu Y. W przeciwnym wypadku stoimy w punkcie X.

Rzecz jasna, generowane w ten sposób zmienne losowe $X(n)$ nie mają dokładnie rozkładu jednostajnego ani tym bardziej nie są niezależne. Jednak dla dużych n zmienna $X(n)$ *ma w przybliżeniu* rozkład jednostajny:

$$\mathbb{P}(X(n) = x) \to \frac{1}{|\mathcal{X}(b)|} \qquad (n \to \infty)$$

dla każdego $x \in \mathcal{X}(b)$. Co więcej, z prawdopodobieństwem 1 jest prawdą, że

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X(i) = x) \to \frac{1}{|\mathcal{X}(b)|} \qquad (n \to \infty).$$

Innymi słowy, ciąg $X(n)$ spełnia prawo wielkich liczb – i może być użyty do szacowania liczności podzbiorów przestrzeni $\mathcal{X}(b)$ *zamiast* trudnego do symulowania ciągu niezależnych zmiennych o rozkładzie jednostajnym.

2. Podstawy R i ćwiczenia komputerowe

Oczekuję, że Czytelnik uruchomił pakiet (lub „otoczenie", *environment*) **R**. Nie zakładam żadnej wstępnej znajomości tego języka programowania. Oczywiście, podstaw programowania w R można się nauczyć z odpowiednich źródeł. Ale prawdę mówiąc, prostych rzeczy można się szybko domyślić i zacząć zabawę natychmiast.

2.1. Początki

Przykład 2.1 (Rozkład prawdopodobieństwa i próbka losowa)**.** Wylosujmy „próbkę" średniego rozmiaru, powiedzmy $n = 100$ z rozkładu normalnego $\mathrm{N}(0, 1)$:

```
> n <- 100
> X <- rnorm(n)
> X
```

(aby przyjrzeć się funkcji `rnorm`, napiszmy `?rnorm`). Wektor `X` zawiera „realizacje" niezależnych zmiennych losowych $X_1, \ldots, X_{100}$. Możemy obliczyć średnią $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ i wariancję próbkową $\bar{X} = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Jakich wyników oczekujemy? Zobaczmy:

```
> mean(X)
> var(X)
```

Porównajmy kwantyl empiryczny rzędu, powiedzmy $p = 1/3$, z teoretycznym:

```
> quantile(X,p)
> qnorm(p)
```

Możemy powtórzyć nasze „doświadczenie losowe" i zobaczyć jakie są losowe fluktuacje wyników. Warto wykonać nasz mini-programik jeszcze raz, albo parę razy. Teraz sprobujmy „zobaczyć" próbkę. Najlepiej tak:

```
> hist(X,prob=TRUE) ( proszę się dowiedzieć co znaczy parametr 'prob' ! )
> rug(X)
```

Możemy łatwo narysować wykres gęstości. Funkcja obliczająca gęstość φ rozkładu $\mathrm{N}(0, 1)$ nazywa się `dnorm`, a `curve` jest funkcją rysującą wykresy.

```
> curve(dnorm(x),col="blue",add=TRUE)
```

(nawiasem mówiąc, zachęcam *początkujących* p**R**obabilistów do oznaczania wygenerowanych wektorów losowych dużymi literami, np. `X`, a deterministycznych zmiennych i wektorów – małymi, np. `x`, podobnie jak na rachunku prawdopodobieństwa).

Podobnie, możemy porównać dystrybuantę empiryczną z prawdziwą dystrybuantą Φ. Funkcja obliczająca dystrybuantę empiryczną nazywa się `ecdf`, zaś dystrybuantę rozkładu $\mathrm{N}(0, 1)$ – `pnorm`.

```
> plot(ecdf(X))
```

```
> curve(pnorm(x),from=xmin,to=xmax,col="blue",add=TRUE)
```

Test Kołmogorowa-Smirnowa oblicza maksymalną odległość $D = \sum_x |\hat{F}_n(x) - F(x)|$, gdzie $\hat{F}_n$ jest dystrybuantą empiryczną i $F = \Phi$.

```
> ks.test(X,pnorm,exact=TRUE)
```

Przypomnijmy sobie z wykładu ze statystyki, co znaczy podana przez test p-wartość. Jakiego wyniku „spodziewaliśmy się"?

Jest teraz dobra okazja, aby pokazać jak się symulacyjnie bada rozkład zmiennej losowej. Przypuśćmy, że interesuje nas rozkład prawdopodobieństwa p-wartości w przeprowadzonym powyżej doświadczeniu (polegającym na wylosowaniu próbki X i przeprowadzeniu testu KS). Tak naprawdę powinniśmy znać odpowiedź bez żadnych doświadczeń, ale możemy udawać niewiedzę i symulować.

Przykład 2.2 (Powtarzanie doświadczenia symulacyjnego)**.** Symulacje polegają na powtórzeniu całego doświadczenia wiele razy, powiedzmy $m = 10000$ razy, zanotowaniu wyników i uznaniu powstałego rozkładu empirycznego za przybliżenie badanego rozkładu prawdopodobieństwa. Bardzo ważne jest zrozumienie różnej roli jaką pełni tu n (rozmiar próbki w pojedynczym doświadczeniu, a więc „parametr badanego zjawiska") i m (liczba powtórzeń, która powinna być możliwie największa aby zwiększyć dokładność badania symulacyjnego). Następujący progarmik podkreśla logiczną konstrukcję powtarzanego doświadczenia:

```
> m <- 10000
> n <- 100
>
> # Przygotowujemy wektory w którym zapiszemy wyniki:
> D <- c()
> P <- c()
>
> for (i in 1:m)
> {
> X <- rnorm(n)
> Test <- ks.test(X,pnorm,exact=TRUE)
> D[i] <- Test$statistic
> P[i] <- Test$p.value
> } # koniec pętli for
>
> # Analizujemy wyniki:
> hist(D, prob=TRUE)
> hist(P, prob=TRUE)
```

Co prawda powyższy programik spełnia swoją rolę, jednak struktura pakiektu R jest dostosowana do innego stylu pisania programów. Zamiast pętli `for` zalecane jest używanie funkcji które powtarzają pewne operacje i posługiwanie się, kiedykolwiek to możliwe, całymi wektorami a nie pojedynczymi komponentami. Poniższy fragment kodu zastępuje pętlę `for`

```
> DiP <- replicate(m, ks.test(rnorm(n),pnorm,exact=TRUE)[1:2]))
>
> DiP <- t(DiP) # transpozycja macierzy ułatwia oglądanie wyników
> D <- as.numeric((DiP)[,1]) # pierwsza kolumna macierzy
> P <- as.numeric((DiP)[,2]) # druga kolumna macierzy
```

```
> # obiekty typu ,,list'' przerobimy na wektory liczbowe:
> D <- as.numeric(D); P <- as.numeric(D)
```

Symulacje dają okazję „namacalnego" przedstawienia twierdzeń probabilistycznych (i nie tylko twierdzeń, ale także stwierdzeń, przypuszczeń prawdziwych lub fałszywych).

Przykład 2.3 (Mocne Prawo Wielkich Liczb)**.** Wylosujmy próbkę z „jakiegoś" rozkładu prawdopodobieństwa. Weźmy na przykład niezależne zmienne o rozkładzie wykładniczym, $X_1, \ldots, X_n, \ldots$ Ex(2). Obliczmy średnie empiryczne

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

Zróbmy wykres ciągu średnich $S_1/1, S_2/2, \ldots, S_n/n, \ldots$.

```
   > nmax <- 1000 # komputerowy odpowiednik ,,n → ∞''
> n <- (1:nmax)
> lambda <- 2 > X <- rexp(nmax,rate=lambda)
> S <- cumsum(X) # ciąg narastających sum
> M <- S/n # działania w R (np. dzielenie) są wykonywane ,,po współrzędnych''
> plot(n,M,type="l") # ,,zaklęcie'' type="l" powoduje narysowanie łamanej
```

Teraz spóbujmy podobne doświadczenie zrobić dla zmiennych $X_i = 1/U_i - 1$, gdzie $U_i \sim$ U$(0,1)$ (X_i są próbką z tak zwanego rozkładu Pareto).

```
   > # potrzebujemy dużej próbki, żeby się zorientować, co się dzieje...
> nmax <- 100000
> n <- (1:nmax)
> X <- 1/runif(nmax)-1
> M <- cumsum(X)/n
> plot(log(n),M,type="l") # i zrobimy wykres w skali logarytmicznej
```

Przykład 2.4 (Centralne Twierdzenie Graniczne)**.** CTG jest jednym ze „słabych" twierdzeń granicznych rachunku prawdopodobieństwa, to znaczy dotyczy zbieżności rozkładów. Symulacyne „sprawdzenie" lub ilustracja takich twierdzeń wymaga powtarzania doświadczenia wiele razy, podobnie jak w Przykładzie 2.2. Pojedyncze doświadczenie polega na obliczeniu sumy

$$S_n = \sum_{i=1}^{n} X_i,$$

gdzie $X_1, \ldots, X_n$ jest próbką z „jakiegoś" rozkładu prawdopodobieństwa i n jest „duże". Weźmy na przykład niezależne zmienne o rozkładzie wykładniczym, jak w Przykładzie 2.3.

```
   > m <- 10000
> n <- 100
> lambda <- 2
> S <- replicate(m, sum(rexp(n,rate=lambda)))
> hist(S,prob=TRUE)
> curve(dnorm(x,mean=n/lambda,sd=sqrt(n)/lambda),col="blue",add=TRUE)
> # wydaje się na podstawie obrazka, że dopasowanie jest znakomite
> ks.test(S,pnorm,mean=n/lambda,sd=sqrt(n)/lambda)
> # ale test Kołmogorowa-Smirnowa ,,widzi'' pewne odchylenie od rozkładu normalnego
>
```

Chciałbym podkreślić raz jeszcze różnicę pomiędzy metodologią sprawdzania „mocnego" twierdzenia granicznego w Przykładzie 2.3 i „słabego" w Przykładzie 2.4. W pierwszym przypadku chcieliśmy zobrazować zbieżność ciągu *zmiennych losowych*, a w drugim – zbieżność ciągu *rozkładów prawdopodobieństw*a.

2.2. Ćwiczenia

Ćwiczenie 2.1. Wyjaśnić dlaczego w Przykładzie 2.4 nie musieliśmy (choć mogliśmy) rozpatrywać *unormowanych* zmiennych losowych

$$\frac{S_n - n\mu}{\sqrt{n}\sigma},$$

gdzie $\mu = \mathbb{E}X_1$ i $\sigma^2 = \mathrm{Var}\, X_1$.

Ćwiczenie 2.2. W Przykładzie 2.4 znany jest *dokładny* rozkład prawdopodobieństwa sumy S_n. Co to za rozkład? Sprawdzić symulacyjnie zgodność z tym rozkładem.

Ćwiczenie 2.3. Przeprowadzić podobne doświadczenie jak w Przykładzie 2.4 (CTG), dla $n = 12$ i $X_i \sim \mathrm{U}(0,1)$. Skomentować wynik.

Ćwiczenie 2.4. Przeprowadzić podobne doświadczenie jak w Ćwiczeniu 2.3, ale zastępując sumę przez medianę dla $n = 13$ i $X_i \sim \mathrm{U}(0,1)$. Wypróbować przybliżenie normalne (jeśli teoretyczna wartość wariancji mediany nie jest znana, można zastąpić ją przez przybliżenie empiryczne). Skomentować wynik.

Ćwiczenie 2.5. W Ćwiczeniu 2.3, znany jest *dokładny* rozkład prawdopodobieństwa mediany. Co to za rozkład? Sprawdzić symulacyjnie zgodność z tym rozkładem. Można wziąć mniejsze, nieparzyste n. Dlaczego nieparzyste?

3. Generowanie zmiennych losowych I. Ogólne metody

3.1. Przykłady

Moje wykłady ograniczają się do zagadnień leżących w kompetencji rachunku prawdopodobieństwa i statystyki. U podstaw symulacji stochastycznych leży generowanie „liczb pseudo-losowych", naśladujących zachowanie zmiennych losowych o rozkładzie jednostajnym. Jak to się robi, co to jest „pseudo-losowość", czym się różni od „prawdziwej losowości"? To są fascynujące pytania, którymi zajmuje się: teoria liczb, teoria (chaotycznych) układów dynamicznych oraz filozofia. Dyskusja na ten temat przekracza ramy tych wykładów. Z punktu widzenia użytkownika, „liczby losowe" są bardzo łatwo dostępne, bo ich generatory są wbudowane w systemy komputerowe. Przyjmę pragmatyczny punkt widzenia i zacznę od następującego założenia.

Założenie 3.1. Mamy do dyspozycji potencjalnie nieskończony ciąg niezależnych zmiennych losowych $U_1, \dots U_n \dots$ o jednakowym rozkładzie $\mathrm{U}(0,1)$.

W języku algorytmicznym: przyjmujemy, że każdorazowe wykonanie instrukcji zapisanej w pseudokodzie

Listing.

```
Gen U ~ U(0,1)
```

wygeneruje kolejną (nową) zmienną U_n. Innymi słowy, zostanie wykonane nowe, niezależne doświadczenie polegające na wylosowaniu przypadkowo wybranej liczby z przedziału $]0,1[$.

Przykład 3.1. Wykonanie pseudokodu

Listing.

```
for i = 1 to 10
  begin
  Gen U ~ U(0,1)
  write U
end
```

da, powiedzmy, taki efekt:

0.32240106 0.38971803 0.35222521 0.22550039 0.04162166
0.13976025 0.16943910 0.69482111 0.28812341 0.58138865

Nawiasem mówiąc, rzeczywisty kod w R, który wyprodukował nasze 10 liczb losowych był taki:

```
U <- runif(10); U
```

Język R jest zwięzły i piękny, ale nasz pseudokod ma pewne walory dydaktyczne.

Zajmiemy się teraz pytaniem, jak „wyprodukować" zmienne losowe o różnych rozkładach, wykorzystując zmienne $U_1, U_2, \ldots$. W tym podrozdziale pokażę kilka przykładów, a w następnym przedstawię rzecz nieco bardziej systematycznie.

Przykład 3.2 (Rozkład Wykładniczy)**.** To jest wyjątkowo łatwy do generowania rozkład – wystarczy taki algorytm:

Listing.

Gen U; $X := -\frac{1}{\lambda}\ln U$

Na wyjściu, $X \sim \mathrm{Ex}(\lambda)$. Żeby się o tym przekonać, wystarczy obliczyć dystrybuantę tej zmiennej losowej: $\mathbb{P}(X \leqslant x) = \mathbb{P}(-\frac{1}{\lambda}\log U \leqslant x) = \mathbb{P}(U \geqslant \mathrm{e}^{-\lambda x}) = 1 - \mathrm{e}^{-\lambda x}$. Jest to najprostszy przykład ogólnej metody „odwracania dystrybuanty", której poświęcę następny podrozdział.

Przykład 3.3 (Generacja rozkładu normalnego)**.** Zmienna losowa

$$X = \sum_{i=1}^{12} U_1 - 6$$

ma *w przybliżeniu* standardowy rozkład normalny $\mathrm{N}(0,1)$. Wynika to z Centralnego Twiedzenia Granicznego (jeśli uznamy, że liczba 12 jest dostatecznie bliska ∞; zauważmy, że $\mathbb{E}X = 0$ i $\mathrm{Var}X = 1$). Oczywiście, w czasach szybkich komputerów ta przybliżona metoda zdecydowanie nie jest polecana. Jest natomiast pouczające zbadać (symulacyjnie!) jak dobre jest przybliżenie. Faktycznie bardzo trudno odróżnić próbkę $X_1, \ldots, X_n$ wyprodukowaną przez powyższy algorytm od próbki pochodzącej *dokładnie* z rozkłdu $\mathrm{N}(0,1)$ (chyba, że n jest ogromne).

Przykład 3.4 (Algorytm Boxa-Müllera)**.** Oto, dla porównania, bardziej współczesna – i całkiem dokładna metoda generowania zmiennych o rozkładzie normalnym.

Listing.

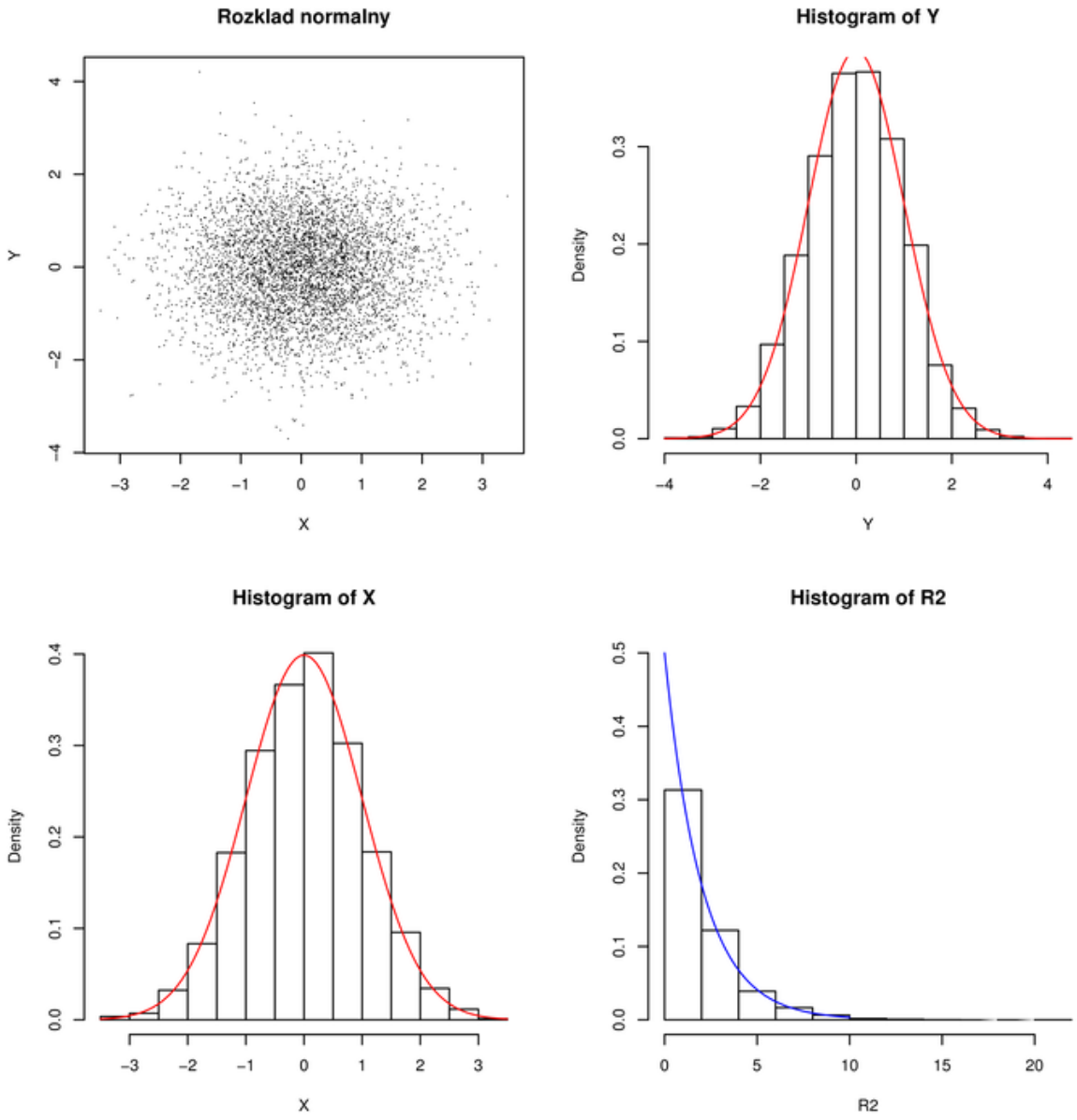
Gen U_1; $\Theta := 2\pi U_1$,
Gen U_2; $R := \sqrt{-2\ln U_2}$,
Gen $X := R\cos\Theta$; $Y := R\sin\Theta$

Na wyjściu *obie* zmienne X i Y mają rozkład $\mathrm{N}(0,1)$ i w dodatku są niezależne. Uzasadnienie poprawności algorytmu Boxa-Müllera opiera się na dwu faktach: zmienna $R^2 = X^2 + Y^2$ ma rozkład $\chi^2(2) = \mathrm{Ex}(1/2)$, zaś kąt Θ między osią i promieniem wodzącym punktu (X, Y) ma rozkład $\mathrm{U}(0, 2\pi)$.

Ciekawe, że łatwiej jest generować zmienne losowe normalne „parami".

Doświadczenie, polegające na wygenerowaniu zmiennych losowych X i Y powtórzyłem 10000 razy. Na Rysunku 3.1 widać 10000 wylosowanych w ten sam sposób i niezależnie punktów (X, Y), histogramy i gęstości brzegowe X i Y (każda ze współrzędnych ma rozkłd $\mathrm{N}(0,1)$) oraz histogram i gęstość $R^2 = X^2 + Y^2$ (rozkład wykładniczy $\mathrm{Ex}(1/2)$).

Histogram jest „empirycznym" (może w obecnym kontekście należałoby powiedzieć „symulacyjnym") odpowiednikiem gęstości: spośród wylosowanych wyników zliczane są punkty należące do poszczególnych przedziałów.



Rysunek 3.1. Dwuwymiarowy rozkład normalny i rozkłady brzegowe.

Przykład 3.5 (Rozkład Poissona)**.** Algorytm przedstawiony niżej nie jest efektywny, ale jest prosty do zrozumienia.

Listing.

$c := \mathrm{e}^{-\lambda}$
Gen U; $P := U$; $N := 0$
while $P \geqslant c$ **do**
  begin Gen U; $P := PU$; $N := N + 1$ **end**

Na wyjściu $N \sim \text{Poiss}(\lambda)$. Istotnie, niech $E_1, \ldots, E_n, \ldots$ będą i.i.d. $\sim \text{Ex}(1)$ i $S_n = E_1 + \cdots + E_n$. Wykorzystamy znany fakt, że $N = \max\{n : S_n \leqslant \lambda\}$ ma rozkład $\text{Poiss}(\lambda)$. Zmienne o rozkładzie wykładniczym przedstawiamy jako $E_i = -\ln U_i$ – patrz Przykład 3.2. Zamiast

dodawać zmienne E_i możemy mnożyć zmienne U_i. Mamy $N = \max\{n : P_n \geqslant \mathrm{e}^{-\lambda}\}$, gdzie $P_n = U_1 \cdots U_n$.

3.2. Metoda przekształceń

Zmienna losowa Y, która ma postać $Y = h(X)$, a więc jest pewną funkcją zmiennej X, w naturalny sposób „dziedziczy" rozkład prawdopodobieństwa zgodnie z ogólnym schematem $\mathbb{P}(Y \in \cdot) = \mathbb{P}(h(X) \in \cdot) = \mathbb{P}(X \in h^{-1}(\cdot))$ („wykropkowany" argument jest zbiorem). Przy tym zmienne X i Y nie muszą być jednowymiarowe. Jeśli obie zmienne mają *ten sam wymiar* i przekształcenie h jest dyfeomorfizmem, to dobrze znany wzór wyraża gęstość rozkładu Y przez gęstość X (Twierdzenie 5.1). Odpowiednio dobierając funkcję h możemy „przetwarzać" jedne rozkłady prawdopodobieństwa na inne, nowe.

Prawie wszystkie algorytmy generowania zmiennych losowych zawierają przekształcenia zmiennych losowych jako część składową. W niemal „czystej" postaci metoda przekształceń pojawiła się w algorytmie Boxa-Müllera, Przykładzie 3.4. Najważniejszym być może szczególnym przypadkiem metody przekształceń jest odwracanie dystrybuanty.

3.2.1. Odwrócenie dystrybuanty

Faktycznie, ta metoda została już wykorzystana w Przykładzie 3.2. Opiera się ona na prostym fakcie. Jeżeli F jest ciągłą i ściśle rosnącą dystrybuantą, $U \sim \mathrm{U}(0,1)$ i $X = F^{-1}(U)$, to $X = F^{-1}(U) \sim F$. Następująca definicja funkcji „pseudo-odwrotnej" pozwala pozbyć się kłopotliwych założeń.

Definicja 3.1. Jeżeli $F : \mathbb{R} \to [0,1]$ jest dowolną dystrybuantą, to funkcję $F^- :]0,1[\to \mathbb{R}$ określamy wzorem:

$$F^-(u) = \inf\{x : F(x) \geqslant u\}.$$

Stwierdzenie 3.1. *Nierówność $F^-(u) \leqslant x$ jest równoważna $u \leqslant F(x)$, dla dowolnych $u \in]0,1[$ i $x \in \mathbb{R}$.*

Dowód. Z prawostronnej ciągłości dystrybuanty F wynika, że kres dolny w Definicji 3.1 jest *osiągany*, czyli

$$F(F^-(u)) \geqslant u.$$

Z drugiej strony,

$$F^-(F(x)) = \min\{y : F(y \geqslant F(x)\} \leqslant x,$$

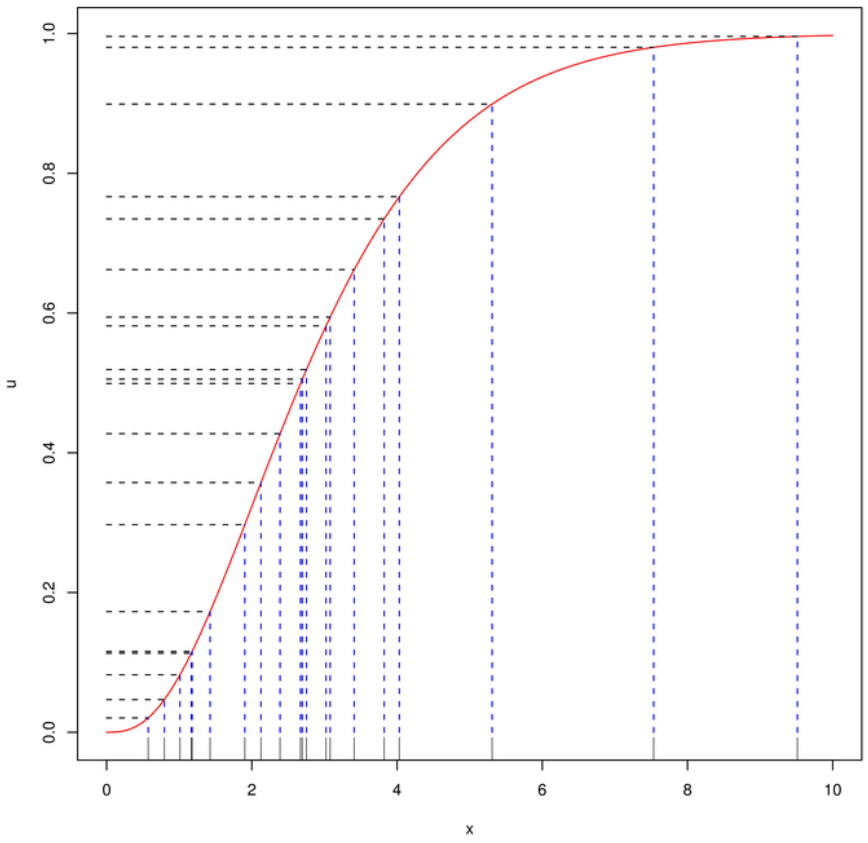
po prostu dlatego, że $x \in \{y : F(y \geqslant F(x)\}$. Teza stwierdzenia natychmiast wynika z dwóch nierówności powyżej. □

Wniosek 3.1 (Ogólna metoda odwrócenia dystrybuanty)**.** *Jeżeli $U \sim \mathrm{U}(0,1)$ i $X = F^-(U)$, to $\mathbb{P}(X \leqslant x) = F(x)$. W skrócie, $X \sim F$.*

Na Rysunku 3.2 widać 20 punktów $U_1, \ldots, U_{20}$, które wylosowałem z rozkładu $\mathrm{U}(0,1)$ (na osi pionowej) i odpowiadające im punkty $X_i = F^{-1}(U_i)$ (na osi poziomej). W tym przypadku, F jest dystrybuantą rozkładu Gamma$(3,1)$ (linia krzywa). Najważniejszy fragment kodu w R jest taki:

```
curve(pgamma(x,shape=3,rate=1), from=0,to=10)   # rysowanie F
U <- runif(20); X <- qgamma(U,shape=3,rate=1)
```

Zauważmy, że ta metoda działa również dla rozkładów dyskretnych i sprowadza się wtedy do metody „oczywistej".



Rysunek 3.2. Odwracanie dystrybuanty.

Przykład 3.6 (Rozkłady dyskretne)**.** Załóżmy, że $\mathbb{P}(X = i) = p_i$ dla $i = 1, 2, \ldots$ i $\sum p_i = 1$. Niech $s_0 = 0$, $s_k = \sum_{i=1}^{k} p_i$. Jeżeli F jest dystrybuantą zmiennej losowej X, to

$$F^-(u) = i \quad \text{wtedy i tylko wtedy gdy} \quad s_{i-1} < u \leqslant s_i.$$

Odwracanie dystrybuanty ma ogromne znaczenie teoretyczne, bo jest całkowicie ogólną metodą generowania dowolnych zmiennych losowych jednowymiarowych. Może się to wydać dziwne, ale w praktyce ta metoda jest używana stosunkowo rzadko, z dwóch względów:
— Obliczanie F^- j bywa trudne i nieefektywne.
— Stosowalność metody ogranicza się do zmiennych losowych jednowymiarowych.
Podam dwa przykłady, w których zastosowanie metody odwracanie dystrybuanty jest rozsądne

Przykład 3.7 (Rozkład Weibulla)**.** Z definicji, $X \sim \text{Weibull}(\beta)$, jeśli

$$F(x) = 1 - \exp(-x^\beta)$$

dla $x \geqslant 0$. Odwrócenie dystrybuanty i generacja X są łatwe:

$$X = (-\ln U)^{1/\beta}, \quad U \sim \text{U}(0, 1).$$

Przykład 3.8 (Rozkład Cauchy'ego)**.** Gęstość i dystrybuanta zmiennej $X \sim \text{Cauchy}(0,1)$ są następujące:

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \qquad F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x).$$

Można tę zmienną generować korzystając z wzoru:

$$X = \tan\left(\pi\left(U - \frac{1}{2}\right)\right), \quad U \sim \text{U}(0,1).$$

3.3. Metoda eliminacji

To jest najważniejsza, najczęściej stosowana i najbardziej uniwersalna metoda. Zacznę od raczej oczywistego faktu, który jest w istocie probabilistycznym sformułowaniem definicji prawdopodobieństwa warunkowego.

Stwierdzenie 3.2. *Przypuśćmy, że $Z = Z_1, \ldots, Z_n, \ldots$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie, o wartościach w przestrzeni $\mathcal{Z}$. Niech $C \subseteq \mathcal{Z}$ będzie takim zbiorem, że $\mathbb{P}(Z \in A) > 0$. Niech*

$$N = \min\{n : Z_n \in C\}.$$

Zmienne losowe N i Z_N są niezależne, przy tym

$$\mathbb{P}(Z_N \in B) = \mathbb{P}(Z \in B | Z \in C) \quad \textit{dla dowolnego} \quad B \subseteq \mathcal{Z},$$

zaś

$$\mathbb{P}(N = n) = pq^{n-1}, \quad (n = 1, 2, \ldots), \quad \textit{gdzie} \quad p = \mathbb{P}(Z \in C).$$

Dowód. Wystarczy zauważyć, że

$$\begin{aligned}\mathbb{P}(X_N \in B, N = n) &= \mathbb{P}(Z_1 \notin C, \ldots, Z_{n-1} \notin C, Z_n \in C \cap B) \\ &= \mathbb{P}(Z_1 \notin C) \cdots \mathbb{P}(Z_{n-1} \notin C)\mathbb{P}(Z_n \in C \cap B) \\ &= q^{n-1}\mathbb{P}(Z \in C \cap B) = q^{n-1}\mathbb{P}(Z \in B | Z \in C)p.\end{aligned}$$

□

W tym Stwierdzeniu $\mathcal{Z}$ może być dowolną przestrzenią mierzalną, zaś C i B – dowolnymi zbiorami mierzalnymi. Stwierdzenie mówi po prostu, że prawdopodobieństwo warunkowe odpowiada doświadczeniu losowemu *powtarzanemu aż do momentu spełnienia warunku*, przy czym rezultaty poprzednich doświadczeń się ignoruje (stąd nazwa: eliminacja).

3.3.1. Ogólny algorytm

Zakładamy, że umiemy generować zmienne losowe o gęstości g, a chcielibyśmy otrzymać zmienną o gęstości *proporcjonalnej* do funkcji f. Zakładamy, że $0 \leqslant f \leqslant g$.

Listing.

repeat
  Gen $Y \sim g$;
  Gen $U \sim \text{U}(0,1)$
until $U \leqslant \dfrac{f(Y)}{g(Y)}$;
$X := Y$

Dowód poprawności algorytmu. Na mocy Stwierdzenia 3.2, zastosowanego do zmiennych losowych $Z = (Y, U)$ wystarczy pokazać, że

$$\mathbb{P}(Y \in B | U \leqslant f(Y)/g(Y)) = \frac{\int_B f(y)\mathrm{d}y}{\int_{\mathcal{X}} f(y)\mathrm{d}y}, \tag{3.1}$$

gdzie $\mathcal{X}$ jest przestrzenią wartości zmiennej losowej Y (i docelowej zmiennej losowej X). Warunkujemy teraz przez wartości zmiennej Y, otrzymując

$$\mathbb{P}(Y \in B, U \leqslant f(Y)/g(Y)) \int_B \mathbb{P}(U \leqslant f(Y)/g(Y) | Y = y) g(y)\mathrm{d}y$$
$$\int_B \frac{f(y)}{g(y)} g(y)\mathrm{d}y = \int_B f(y)\mathrm{d}y.$$

Skorzystaliśmy z niezależności Y i U oraz ze wzoru na prawdopodobieństwo całkowite. Otrzymaliśmy *licznik* we wzorze (3.1). *Mianownik* dostaniemy zastępując B przez $\mathcal{X}$. □

Uwaga 3.1. Ważną zaletą algorytmu jest to, że nie trzeba znać „stałej normującej" gęstości $f/\int f$, czyli liczby $\int f$.

Uwaga 3.2. W istocie $\mathcal{X}$ może być ogólną przestrzenią z miarą μ. Żeby poczuć się pewniej założę, że $\mathcal{X}$ jest przestrzenią polską i miara μ jest σ-skończona. Możemy rozważać gęstości i całki względem miary μ – dowód poprawości algorytmu eliminacji nie ulega zmianie. Nie widzę wielkiego sensu w przeładowaniu notacji, można się umówić, że symbol $\int \cdots \mathrm{d}x$ jest *skrótem* $\int \cdots \mu(\mathrm{d}x)$. Dociekliwy Czytelnik powinien zastanowić się, w jakiej sytuacji potrafi uzasadnić wszystkie przejścia w dowodzie, powołując się na odpowiednie własności prawdopodobieństwa warunkowego (np. twierdzenie o prawdopodobieństwie całkowitym itp.). W każdym razie, algorytm eliminacji działa poprawnie, gdy

— $\mathcal{X} = \mathbb{R}^d$, gęstości są względem miary Lebesgue'a;
— $\mathcal{X}$ dyskretna, gęstości są względem miary liczącej.

Uwaga 3.3. Efektywność algorytmu zależy od dobrania gęstości g tak, aby majoryzowała funkcję f ale nie była dużo od niej większa. Istotnie, liczba prób N do zaakceptowania $X := Y$ ma rozkład geometryczny z prawdopodobieństwem sukcesu $p = \int f / \int g$, zgodnie ze Stwierdzeniem 3.2, zatem $\mathbb{E}N = 1/p$. Iloraz p powinien być możliwie bliski jedynki, co jest możliwe jeśli „kształt funkcji g jest podobny do f.

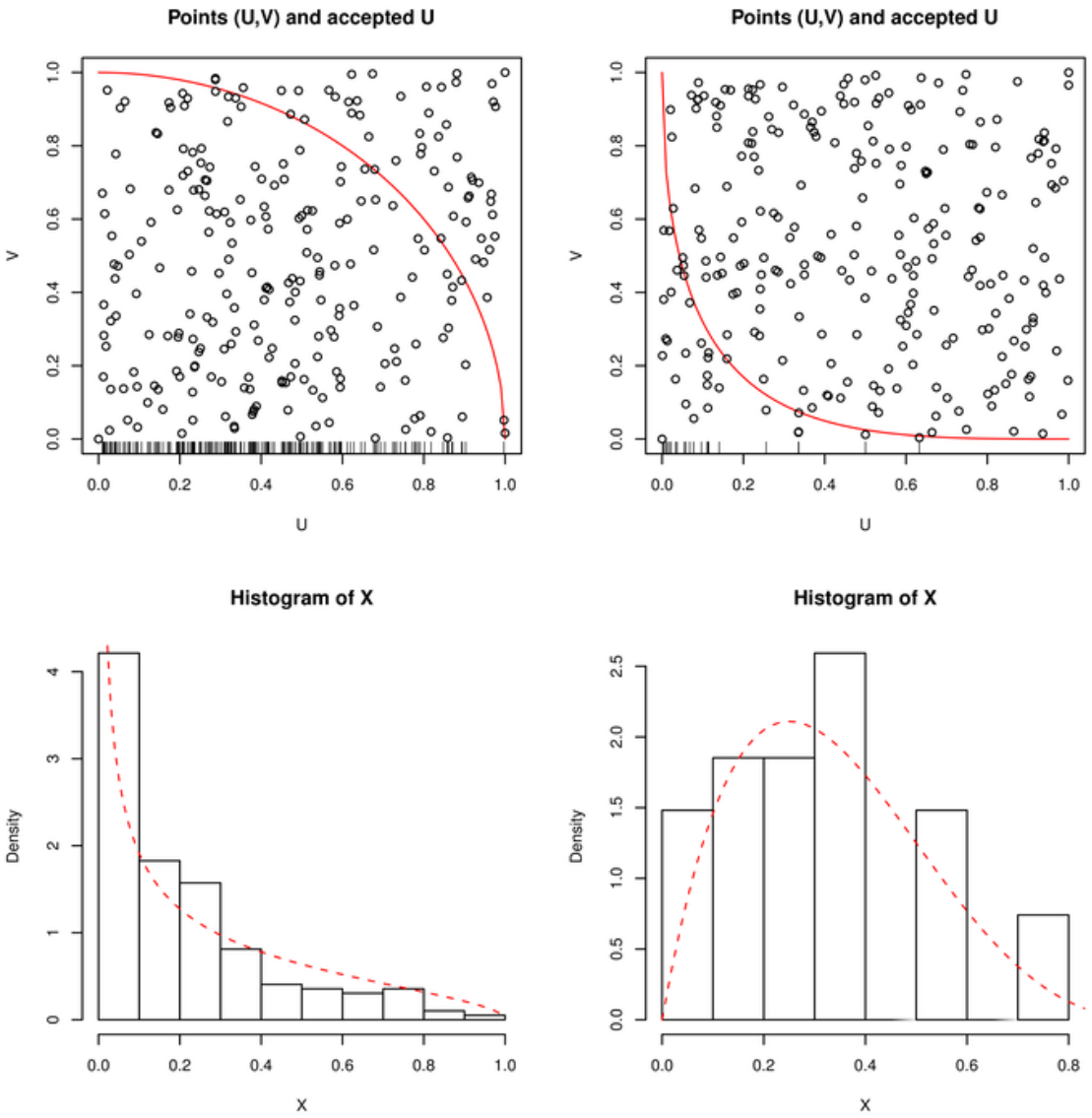
Zwykle metoda eliminacji stosuje się w połączeniu z odpowiednio dobranymi przekształceniami. Zilustrujmy to na poniższym przykładzie.

Przykład 3.9 (Rozkład Beta, algorytm Bermana)**.** Niech U i V będą niezależnymi zmiennymi losowymi o rozkładzie jednostajnym U$(0, 1)$. Pod warunkiem, że $U^{1/\alpha} + V^{1/\beta}$ zmienna losowa U ma gęstość $f_U(u)$ proporcjonalną do funkcji $(1 - u^{1/\alpha})^\beta$. Zmienna losowa $X = U^{1/\alpha}$ ma gęstość $f_X(x) = f_U(x^\alpha)(\mathrm{d}x^\alpha/\mathrm{d}x) \propto (1 - x)^\beta x^\alpha$. Zatem X ma rozkład Beta$(\alpha, \beta + 1)$. Algorytm jest więc następujący:

Listing.

repeat
  Gen U, V
  $X := U^{1/\alpha}$; $Y := V^{1/\beta}$
until $X + Y \leqslant 1$;
return X

Efektywność tego algorytmu jest tym większa, im mniejsze są parametry α i β (frakcja zakceptowanych U jest wtedy duża). Dla ilustracji rozpatrzmy dwa przypadki.



Rysunek 3.3. Algorytm Bermana: losowanie z rozkładu beta.

Na rysunku 3.3, po lewej stronie $\alpha = \beta = 0.5$. Spośród $n = 250$ punktów zaakceptowano $N = 197$. Przerywana krzywa jest wykresem gęstości rozkładu Beta$(0.5, 1, 5)$. Po prawej stronie $\alpha = 2$, $\beta = 3$. Spośród $n = 250$ punktów zaakceptowano tylko $N = 24$. Tym razem wykres pokazuje gęstość rozkładu Beta$(2, 4)$. *Zaakceptowane* punkty U (o gęstości f_U) są widoczne w postaci „kreseczek" na górnych rysunkach. Ciągłą linią jest narysowana funkcja $(1 - u^{1/\alpha})^\beta \propto f_U(u)$.

3.3.2. Eliminacja w R

Zasadniczy algorytm eliminacji opisany w Stwierdzeniu 3.2 i w Punkcie 3.3.1 polega na powtarzaniu generacji tak długo, aż zostanie spełnione kryterium akceptacji. Liczba prób jest losowa, a rezultatem jest jedna zmienna losowa o zadanym rozkładzie. Specyfika języka R narzuca inny sposób przeprowadzania eliminacji. Działamy na wektorach, a więc od razu produkujemy

n niezależnych zmiennych $X_1, \ldots, X_n$ o rozkładzie g, następnie poddajemy je wszystkie procedurze eliminacji. Przez sito eliminacji, powiedzmy $U_i < f(X_i)/g(X_i)$, przechodzi pewna część zmiennych. Otrzymujemy *losową* liczbę zmiennych o rozkładzie proporcjonalnym do f. Liczba zaakceptowanych zmiennych ma oczywiście rozkład dwumianowy $\mathrm{Bin}(n, p)$ z $p = \int f / \int g$. To nie jest zbyt eleganckie. W sztuczny sposób można zmusić R do wyprodukowania *zadanej*, nielosowej liczby zaakceptowanych zmiennych, ale jeśli nie ma specjalnej konieczności, lepiej tego nie robić (przymus rzadko prowadzi do pozytywnych rezultatów).

Dla przykładu, generowanie zmiennych (X, Y) o rozkładzie jesnostajnym na kole $\{x^2 + y^2 \leqslant 1\}$ może w praktyce wyglądać tak:

```
> n <- 1000
> X <- runif(n,min=-1,max=1)  \# generowanie z rozkładu jednostajnego na [-1,1]
> Y <- runif(n,min=-1,max=1)
> Accept <- X^2+Y^2<1    \# wektor logiczny
> X <- X[Accept]
> Y <- Y[Accept]
```

Otrzymujemy pewną losową liczbę (około 785) punktów (X, Y) w kole jednostkowym.

3.4. Metoda kompozycji

Jest to niezwykle prosta technika generowania zmiennych losowych. Załóżmy, że docelowy rozkład jest mieszanką prostszych rozkładów prawdopodobieństwa, czyli jego gęstość jest kombinacją wypukłą postaci

$$f(x) = \sum_{i=1}^{k} p_i f_i(x), \qquad \left(p_i \geqslant 0, \sum_{i=1}^{k} p_i = 1 \right).$$

Jeśli umiemy losować z każdej gęstości f_i to możemy uruchomić dwuetapowe losowanie:

Listing.

Gen $I \sim p(\cdot)$; { *to znaczy* $\mathbb{P}(I = i) = p_i$ }
Gen $X \sim f_I$; { *jeśli* $i = i$ *to uruchamiamy generator rozkładu* f_i }
return X

Jasne, że na wyjściu mamy $X \sim F$. W istocie jest to szczególny przypadek metody rozkładów warunkowych, którą omówię później, w Rozdziale 5.

Przykład 3.10 (Rozkład Laplace'a)**.** Rozkład Laplace'a (podwójny rozkład wykładniczy) ma gęstość

$$f(x) = \frac{1}{2\lambda} \mathrm{e}^{-\lambda |x|}.$$

Można go „skomponować" z dwóch połówek rozkładu wykładniczego:

Listing.

Gen $W \sim \mathrm{Ex}(\lambda)$; { *generujemy z rozkładu wykładniczego* }
Gen $U \sim \mathrm{U}(0, 1)$;
if $U < 1/2$ **then** $X := W$ **else** $X := -W$ { *losowo zmieniamy znak* }
return X

4. Generowanie zmiennych losowych II. Specjalne metody

4.1. Rozkłady dyskretne

Jak wylosować zmienną losową I o rozkładzie $\mathbb{P}(I = i) = p_i$ mając dane $p_1, p_2, \ldots$? Metoda odwracanie dystrybuanty w przypadku rozkładów dyskretnych (Przykład 3.6) przyjmuje następującą postać. Obliczamy $s_1, s_2, \ldots$, gdzie $s_k = \sum_{i=1}^{k} p_i$. Losujemy $U \sim \mathrm{U}(0,1)$ i szukamy przedziału $]s_{I-1}, s_I]$ w którym leży U.

Przykład 4.1 (Algorytm prymitywny)**.** W zasadzie można to zrobić tak:

Listing.

```
  Gen U, I := 1
  while s_I ⩽ U do I := I + 1;
return I
```

Problemem jest mała efektywność tego algorytmu. Jedno z możliwych ulepszeń polega na bardziej „inteligentnym" lokalizowaniu przedziału $]s_{I-1}, s_I] \ni U$, na przykład metodą bisekcji.

Przykład 4.2 (Algorytm podziału odcinka)**.** Załóżmy, że mamy rozkład prawdopodobieństwa (p_i) na przestrzeni skończonej i liczby p_i są uporządkowane: $p_1 > \ldots > p_m$.

Listing.

```
Gen U;
L := 0; R := m;
repeat
  I := ⌊(L + R)/2⌋;
  if U > s_I then L := I else R := I;
until L ⩾ R − 1
return I
```

Przedstawię teraz piękną metodę opartą na innym pomyśle.

Przykład 4.3 (Metoda „Alias")**.** Przypuśćmy, że mamy dwa ciągi liczb: $q_1, \ldots, q_m$, gdzie $0 < q_i < 1$ oraz $a(1), \ldots, a(m)$, gdzie $a(i) \in \{1, \ldots, m\}$ (to są owe „aliasy"). Rozaptrzmy taki algorytm:

Listing.

```
Gen K ∼ U{1, ..., M};
Gen U;
if U < q_K then I := K else I := a(K);
return I
```

Jest jasne, że

$$\mathbb{P}(I=i)=\frac{1}{m}\left(q_i+\sum_{j:a(j)=i}(1-q_j)\right).$$

Jeśli mamy zadany rozkład prawdopodobieństwa $(p_1,\ldots,p_m)$ i chcemy, żeby $\mathbb{P}(I=i)=p_i$, to musimy dobrać odpowiednio q_i i $a(i)$. Można zawsze tak zrobić, i to na wiele różnych sposobów. Opracowanie algorytmu dobierania wektorów q i a do zadanego p pozostawiamy jako ćwiczenie.

4.2. Schematy kombinatoryczne

4.2.1. Pobieranie próbki bez zwracania

Spośród r obiektów chcemy wybrać losowo n tak, aby każdy z $\binom{r}{n}$ podzbiorów miał jednakowe prawdopodobieństwo. Oczywiście, można losować tak, jak ze zwracaniem, tylko odrzucać elementy wylosowane powtórnie.

Przykład 4.4 (Losowanie bez zwracania I)**.** W tablicy $c(1),\ldots,c(r)$ zaznaczamy, które elementy zostały wybrane.

Listing.

```
for i := 1 to r do c(i) := false;
i := 0
repeat
  repeat
    Gen K ~ U{1,...,r}
  until c(K) = false;
  c(K) := true;
  i := i + 1;
until i = n
```

Pewnym ulepszeniem tej prymitywnej metody jest następujący algorytm.

Przykład 4.5 (Losowanie bez zwracania II)**.** Tablica $c(1),\ldots,c(r)$ ma takie samo znaczenie jak w poprzednim przykładzie. Będziemy teraz „przeglądać" elementy $1,\ldots,r$ po kolei, decydując o zaliczeniu do próbki kolejnego elementu zgodnie z odpowiednim prawdopodobieństwem *warunkowym*. Niech i oznacza liczbę wybranych, zaś $t-1$ - liczbę przejrzanych poprzednio elementów.

Listing.

```
for i := 1 to r c(i) := false;
t := 1; i := 0
repeat
  Gen U;
  if U ⩽ (n - i)/(r - (t - 1)) then
    begin
    c(t) := true;
    i := i + 1
  end;
  t := t + 1;
until i = n
```

Przykład 4.6 (Losowanie bez zwracania III)**.** Tablica $s(1), \ldots, s(n)$ będzie teraz zawierała *numery* wybieranych elementów. Podobnie jak poprzednio, i – oznacza liczbę elementów wybranych, $t-1$ – liczbę przejrzanych.

Listing.

```
for i := 1 to n do s(i) := i;
for t := n + 1 to r do
  begin
  Gen U;
  if U ⩽ n/t then
    begin
    Gen K ∼ U{1, ..., n};
    s(K) := t
    end
  end
```

Uzasadnienie poprawności tego algorytmu jest rekurencyjne. Załóżmy, że przed t-tym losowaniem każda próbka wybrana ze zbioru $\{1, \ldots, t-1\}$ ma prawdopodobieństwo

$$\binom{t-1}{n}^{-1} = \frac{n!}{(t-1)\cdots(t-n)}.$$

W kroku t ta próbka „przeżywa" czyli pozostaje bez zmiany z prawdopodobieństwem $1 - n/t$ (jest to prawdopodobieństwo, że próbka wylosowana ze zbioru $\{1, \ldots, t\}$ nie zawiera elementu t. Zatem po kroku t każda próbka nie zawierająca elementu t ma prawdopodobieństwo

$$\frac{n!}{(t-1)\cdots(t-n)} \cdot \frac{t-n}{t} = \binom{t}{n}^{-1}.$$

Z prawdopodobieństwem n/t „podmieniamy" jeden z elementów próbki (losowo wybrany) na element t. Sprawdzenie, że każda probka zawierająca element t ma po t-tym losowaniu jednakowe prawdopodobieństwo – pozostawiam jako ćwiczenie.

4.2.2. Permutacje losowe

Przez permutację losową rozumiemy uporządkowanie n elementów wygenerowane zgodnie z rozkładem jednostajnym na przestrzeni wszystkich $n!$ możliwych uporządkowań. Permutację liczb $1, \ldots, n$ zapiszemy w tablicy $\sigma(1), \ldots, \sigma(n)$. Latwo sprawdzić poprawność następującego algorytmu.

Listing.

```
for i := 1 to n do σ(i) := i;
for i := 1 to n − 1 do
  begin
  Gen J ∼ U{i, i + 1, ..., n};
  Swap(π(i), π(J))
  end
```

Funkcja *Swap* zamienia miejscami elementy $\pi(i)$ i $\pi(J)$.

4.3. Specjalne metody eliminacji

4.3.1. Iloraz zmiennych równomiernych

Szczególnie często stosowany jest specjalny przypadek metody eliminacji, znany jako algorytm „ilorazu zmiennych równomiernych" (*Ratio of Uniforms*). Zaczniemy od prostego przykładu, który wyjaśni tę nazwę.

Przykład 4.7 (Rozkład Cauchy'ego)**.** Metodą eliminacji z prostokąta $[0,1]\times[-1,1]$ otrzymujemy zmienną losową (U,V) o rozkładzie jednostajnym na półkolu $\{(u,v): u\geqslant 0, u^2+v^2\leqslant 1\}$. Kąt Φ pomiędzy osią poziomą i punktem (U,V) ma, oczywiście, rozkład $\mathrm{U}(-\pi,\pi)$.

Listing.

repeat
  Gen $U\sim \mathrm{U}(0,1)$;
  Gen $V\sim \mathrm{U}(-1,1)$
until $U^2+V^2<1$
$X:=V/U$

Na wyjściu $X\sim$ Cauchy, bo

$$\mathbb{P}(X\leqslant x)=\mathbb{P}(V\leqslant xU)=\frac{1}{\pi}\Phi+\frac{1}{2}=\frac{1}{\pi}\arctan(x)+\frac{1}{2}.$$

Ogólny algorytm metody „ilorazu równomiernych" oparty jest na następującym fakcie.

Stwierdzenie 4.1. *Załóżmy o funkcji $h:\mathbb{R}\to\mathbb{R}$, że*

$$h(x)\geqslant 0,\ \int h(x)dx<\infty.$$

Niech zbiór $C_h\subset\mathbb{R}^2$ będzie określony następująco:

$$C_n=\left\{(u,v): 0\leqslant u\leqslant\sqrt{h\left(\frac{u}{v}\right)}\right\},\ |C_n|<\infty.$$

Miara Lebesgue'a (pole) tego zbioru jest skończone, $|C_h|<\infty$, a zatem można mówić o rozkładzie jednostajnym $\mathrm{U}(C_h)$.

Jeżeli $(U,V)\sim\mathrm{U}(C_h)$ i $X=V/U$, to X ma gęstość proporcjonalną do funkcji h ($X\sim h/\int h$).

Dowód. „Pole figury" C_h jest równe

$$|C_h|=\iint\limits_{C_h}\mathrm{d}u\mathrm{d}v=\iint\limits_{0\leqslant u\leqslant\sqrt{h(x)}}u\mathrm{d}u\mathrm{d}x$$
$$=\int\limits_{-\infty}^{\infty}\int\limits_{0}^{\sqrt{h(x)}}\mathrm{d}u\mathrm{d}x=\frac{1}{2}\int\limits_{-\infty}^{\infty}h(x)\mathrm{d}x<\infty$$

Dokonaliśmy tu zamiany zmiennych:

$$(u,v)\mapsto\left(u,x=\frac{v}{u}\right).$$

Jakobian tego przekształcenia jest równy

$$\frac{\partial(u,x)}{\partial(u,v)} = \det\begin{pmatrix} 1 & 0 \\ v/u^2 & 1/u \end{pmatrix} = \frac{1}{u}$$

W podobny sposób, na mocy znanego wzoru na gęstość przekształconych zmiennych losowych obliczamy łączną gęstość (U,X):

$$f_{U,X}(u,x) = u f_{U,V}(u,ux) \quad \text{na zbiorze } \{0 \leqslant u \leqslant \sqrt{h(x)}\}.$$

Stąd dostajemy gęstość brzegową X:

$$\int_0^{\sqrt{h(x)}} \frac{u\,du}{|C_h|} = \frac{h(x)}{2|C_h|}.$$

□

Żeby wylosować $(U,V) \sim \mathrm{U}(C_h)$ stosuje się zazwyczaj eliminację z prostokąta. Użyteczne jest następujące oszacowanie boków tego prostokąta.

Stwierdzenie 4.2. *Jeśli funkcje $h(x)$ i $x^2h(x)$ są ograniczone, wtedy*

$$C_h \subseteq [0,a] \times [b_-, b_+],$$

gdzie

$$a = \sqrt{\sup_x h(x)},$$

$$b_+ = \sqrt{\sup_{x \geqslant 0}[x^2 h(x)]}, \qquad b_- = -\sqrt{\sup_{x \leqslant 0}[x^2 h(x)]}.$$

Dowód. Jeśli $(u,v) \in C_h$ to oczywiście $0 \leqslant u \leqslant \sqrt{h(v/u)} \leqslant \sqrt{\sup_x h(x)}$. Załóżmy dodatkowo, że $v \geqslant 0$ i przejdźmy do zmiennych (v,x), gdzie $x = v/u$. Nierówność $u \leqslant \sqrt{h(v/u)}$ jest równoważna $v^2 \leqslant x^2 h(x)$. Ponieważ $x \geqslant 0$, więc dostajemy $v^2 \leqslant b_+^2$. Dla $v \leqslant 0$ mamy analogicznie $v^2 \leqslant b_-^2$. □

Ogólny algorytm RU jest następujący:

Listing.

repeat
  Gen U_1, U_2;
  $U := aU_1$; $V := b_- + (b_+ - b_-)U_2$
until $(U,V) \in C_h$;
$X := \dfrac{V}{U}$

Przykład 4.8 (Rozkład normalny)**.** Niech

$$h(x) = \exp\left(-\frac{1}{2}x^2\right).$$

Wtedy

$$C_h = \left\{(u,v) : 0 \leqslant u \leqslant \exp\left(-\frac{1}{4}\frac{v^2}{u^2}\right)\right\} = \left\{\frac{v^2}{u^2} \leqslant -4\ln u\right\}.$$

Zauważmy, że $a = 1$ i $b_+ = -b_- = \sqrt{2e^{-1}}$. Otrzymujemy następujący algorytm:

Listing.

repeat
  Gen U_1, U_2
  $U := U_1$; $V := \sqrt{2\mathrm{e}^{-1}}(2U_2 - 1)$;
  $X := \dfrac{V}{U}$
until $X^2 \leqslant -4 \ln U$

4.3.2. Gęstości przedstawione szeregami

Ciekawe, że można skonstruować dokładne algorytmy eliminacji bez konieczności dokładnego obliczania docelowej gęstości f. Podstawowy pomysł jest następujący. Niech f, podobnie jak poprzednio, będzie funkcją proporcjonalną do gęstości (nie jest potrzebna stała normująca) i $f \leqslant g$. Załóżmy, że mamy dwa ciągi funkcji, przybliżające f z dołu i z góry:

$$\underline{f_n} \leqslant f \leqslant \overline{f_n}, \qquad \underline{f_n} \to f,\ \overline{f_n} \to f \quad (n \to \infty).$$

Jeśli umiemy ewaluować funkcje $\underline{f_n}$ i $\overline{f_n}$ to możemy uruchomić następujący algorytm:

Listing.

repeat
Gen $Y \sim g$;
Gen U;
$W := Ug(Y)$;
  repeat
  $n := n + 1$;
  if $W \leqslant \underline{f_n}(Y)$ **then**
    begin $X := Y$; **return** X; stop **end**
  until $W > \overline{f_n}(Y)$
until FALSE

Zbieżność przybliżeń dolnych i górnych do funkcji f gwarantuje, że ten algorytm na pewno się kiedyś zatrzyma. Fakt, że na wyjściu $X \sim f / \int f$ jest widoczny.

Poniższy algorytm jest znany jako **metoda szeregów zbieżnych**. Zakładamy, że

$$f(x) = \sum_{i=1}^{\infty} a_n(x),$$

przy czym reszty tego szeregu umiemy oszacować z góry przez znane funkcje,

$$\left| \sum_{i=n+1}^{\infty} a_n(x) \right| \leqslant r_{n+1}(x).$$

Pseudo-kod algorytmu jest taki:

Listing.

repeat
Gen $Y \sim g$;
Gen U;
$W := Ug(Y)$;
$s := 0$;
$n := 0$;

repeat
$n := n + 1$;
$s := s + a_n(Y)$;
until $|s - W| > r_{n+1}(Y)$;
until $W \leqslant s$;
return X

Rzecz jasna, w tym algorytmie zmienna s przybiera kolejno wartości równe sumom częściowym szeregu, $s_n(x) = \sum_{i=1}^{n} a_i(x)$.

Metoda szeregów naprzemiennych jest dostosowana do sytuacji gdy

$$f(x) = g(x) \sum_{i=0}^{\infty} (-1)^n a_n(x) = g(x) \left[1 - a_1(x) + a_2(x) - a_3(x) \ldots\right],$$

gdzie $1 = a_0(x) \geqslant a_1(x) \geqslant a_2(x) \geqslant \cdots \geqslant 0$. Wiadomo, że w takiej sytuacji sumy częściowe przybliżają sumę szeregu na przemian z nadmiarem i z niedomiarem. Ten fakt wykorzystuje algorytm szeregów naprzemiennych:

Listing.

repeat
Gen $Y \sim g$;
Gen U;
$W := Ug(Y)$;
$s := 0$;
$n := 0$;
repeat
$n := n + 1$; { n *nieparzyste* }
$s := s + a_n(Y)$;
if $U \geqslant s$ **then begin** $X := Y$; **return** X; stop **end**
$n := n + 1$; { n *parzyste* }
$s := s - a_n(Y)$;
until $U > s$;
until FALSE

Przykład 4.9. Rozkład Kołmogorowa-Smirnowa ma dystrybuantę

$$F(x) = \sum_{n=-\infty}^{\infty} (-1)^n \mathrm{e}^{-2n^2x^2}, \quad (x \geqslant 0)$$

i gęstość

$$f(x) = 8 \sum_{n=1}^{\infty} (-1)^{n+1} n^2 x \mathrm{e}^{-2n^2x^2}, \quad (x \geqslant 0).$$

Możemy zastosować metodę szeregów naprzemiennych, kładąc

$$g(x) = 4x\mathrm{e}^{-2x^2} \quad (x \geqslant 0)$$

oraz

$$a_n(x) = n^2 x \mathrm{e}^{-2(n^2-1)x^2}.$$

5. Generowanie zmiennych losowych III. Rozkłady wielowymiarowe

5.1. Ogólne metody

Wiele spośród ogólnych metod generowania zmiennych losowych jest całkowicie niezależnych od wymiaru. W szczególności, metody *eliminacji*, *kompozycji* i *przekształceń* są z powodzeniem stosowane do generowania zmiennych losowych wielowymiarowych. Wyjątek stanowi „najbardziej ogólna" metoda *odwracania dystrybuanty*, która nie ma naturalnego odpowiednika dla wymiaru większego niż 1.

5.1.1. Metoda rozkładów warunkowych

Jest to właściwie jedyna metoda „w zasadniczy sposób wielowymiarowa". Opiera się na przedstawieniu gęstości łącznej zmiennych losowych $X_1, \ldots, X_d$ jako iloczynu gęstości brzegowej i gęstości warunkowych (wzór łańcuchowy):

$$f(x_1, x_2, \ldots, x_d) = f(x_1) f(x_2|x_1) f(x_3|x_1, x_2) \cdots f(x_d|x_1, \ldots, x_{d-1}).$$

Wynika stąd następujący algorytm:

Listing.

for $i := 1$ **to** d **do**
  Gen $X_i \sim f(\cdot|X_1, \ldots, X_{i-1})$

Przykład 5.1 (Wielowymiarowy rozkład normalny)**.** Ograniczymy się do zmiennych losowych X_1, X_2 pochodzących z rozkładu dwuwymiarowego $\mathrm{N}(0, 0, \sigma_1^2, \sigma_2^2, \rho)$ o gęstości

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\Big[-\frac{1}{2(1-\rho^2)}\Big(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\Big)\Big].$$

Jak wiadomo (można to sprawdzić elementarnym rachunkiem),

$$f(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{x_1^2}{2\sigma_1^2} \right],$$

$$f(x_2|x_1) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\Big[-\frac{1}{2\sigma_2^2(1-\rho^2)}\Big(x_2 - \rho\frac{\sigma_2}{\sigma_1}x_1\Big)^2\Big].$$

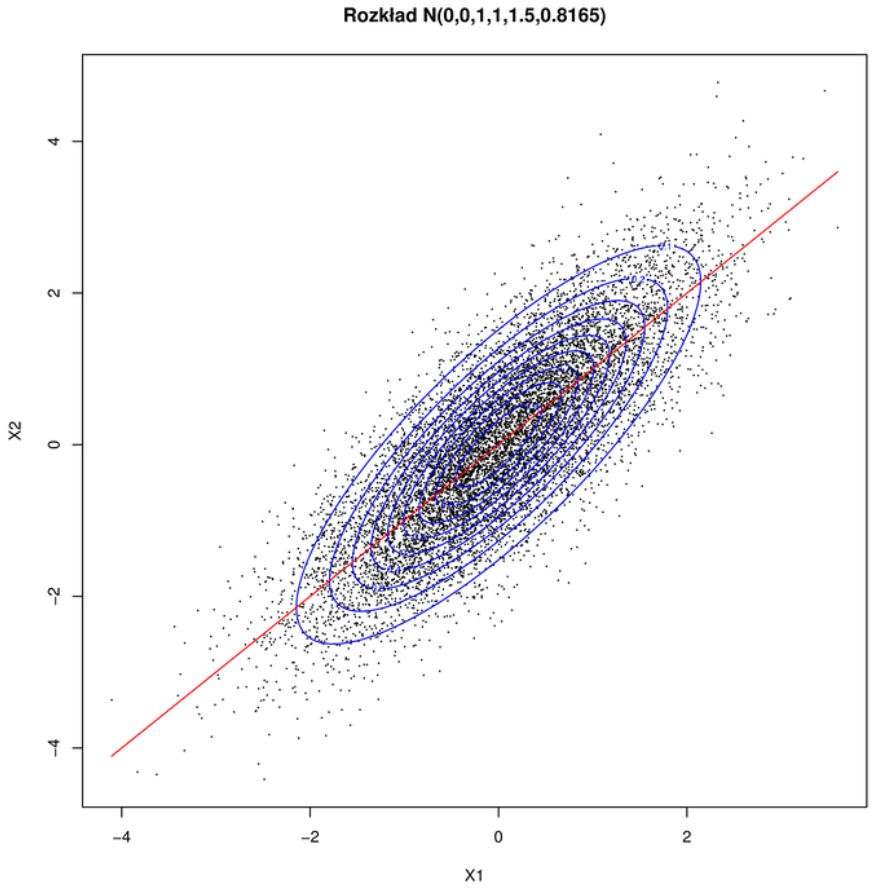
To znaczy, że $\mathrm{N}(0, \sigma_1^2)$ jest rozkładem brzegowym X_1 oraz

$$\mathrm{N}(\rho\frac{\sigma_2}{\sigma_1}x_1, \sigma_2^2(1-\rho^2))$$

jest rozkładem warunkowym X_2 dla $X_1 = x_1$. Algorytm jst więc następujący.

Listing.

Gen $X_1, X_2 \sim \mathrm{N}(0,1)$
$X_1 := \sigma_1 X_1$
$X_2 := \rho(\sigma_2/\sigma_1)X_1 + \sigma_2\sqrt{1-\rho^2}X_2$



Rysunek 5.1. Próbka z dwuwymiarowego rozkładu normalnego, poziomice gęstości i funkcja regresji $x_2 = \mathbb{E}(X_2|X_1 = x_1)$.

Efekt działania powyższego algorytmu widać na Rysunku 5.1. W tym konkretnym przykładzie X_1 ma rozkład brzegowy $\mathrm{N}(0,1)$, zaś X_2 ma rozkład warunkowy $\mathrm{N}(X_1, 0.5)$. Zauważmy, że $\mathrm{Var}X_2 = 1.5$ i $\mathrm{Cov}(X_1, X_2) = 1$. Prosta $x_2 = x_1$ jest wykresem funkcji regresji $\mathbb{E}(X_2|X_1 = x_1)$ i jest przedstawiona na wykresie. Pokazane są też poziomice gęstości (elipsy). Warto zwrócić uwagę, że funkcja regresji nie pokrywa się ze wspólną osią tych elips. Dlaczego? Jak obliczyć oś?

Uogólnienie na przypadek rozkładu normalnego $\mathrm{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, z niezerowymi średnimi, jest banalne. Uogólnienie na wyższe wymiary też nie jest skomplikowane. Okazuje się jednak, że metoda rozkładów warunkowych dla rozkładów normalnych prowadzi to do algorytmu identycznego jak otrzymany metodą przekształceń.

5.1.2. Metoda przekształceń

Podstawą jest następujące twierdzenie o przekształcaniu gęstości.

Twierdzenie 5.1. *Załóżmy, że X jest d-wymiarową zmienną losową o wartościach w otwartym zbiorze $A \subseteq \mathbb{R}^d$. Jeżeli*

$$X \sim f(x)$$

i $h : A \to \mathbb{R}^d$ jest dyfeomorfizmem, to

$$Y = h(X) \sim g(y) = f(h^{-1}(y)) \left| \det \frac{\partial}{\partial y} h^{-1}(y) \right|.$$

Przykład 5.2 (Wielowymiarowy rozkład normalny)**.** Rozważmy niezależne zmienne losowe $Z_1, \ldots, Z_d \sim \mathrm{N}(0,1)$. Wektor $Z = (Z_1, \ldots, Z_d)$ ma d-wymiarowy rozkład normalny $\mathrm{N}(0, I)$ o gęstości

$$f(z) = (2\pi)^{-d/2} \exp\left[-\frac{1}{2} z^\top z\right].$$

Jeżeli teraz R jest nieosobliwą macierzą $(d \times d)$ to przekształcenie $z \mapsto x = Rz$ jest dyfeomorfizmem z jakobianem $\det \mathrm{R}$. Z Twierdzenia 5.1 wynika, że wektor losowy $X = \mathrm{R}Z$ ma rozkład normalny o gęstości

$$f_X(x) = (2\pi)^{-d/2} (\det \mathrm{R})^{-1/2} \exp\left[-\frac{1}{2} x^\top \Sigma^{-1} x\right],$$

gdzie $\Sigma = \mathrm{R}\mathrm{R}^\top$. Innymi słowy, $X \sim \mathrm{N}(0, \Sigma)$. Algorytm generacji jest oczywisty:

Listing.

Gen $Z \sim \mathrm{N}(0, I)$
$X := \mathrm{R}Z$

Jeśli dana jest macierz kowariancji Σ wektora X, to przed uruchomieniem algorytmu trzeba znaleźć taką macierz R, żeby $\Sigma = \mathrm{R}\mathrm{R}^\top$. Istnieje wiele takich macierzy, ale najlepiej skorzystać z rozkładu Choleskiego i wybrać macirz trójkątną.

5.2. Kilka ważnych przykładów

Rozmaitych rozkładów wielowymiarowych jest więcej, niż gwiazd na niebie – a wśród nich tyle przykładów ciekawych i ważnych! Musiałem wybrać zaledwie kilka z nich. Rzecz jasna, wybrałem te, które mi się szczególnie podobają.

5.2.1. Rozkłady sferyczne i eliptyczne

Najważniejszy, być może, przykład to wielowymiarowy rozkład normalny, omówiony już w 5.2. Rozpatrzymy teraz *rozklady jednostajne* na *kuli*

$$B^d = \{x \in \mathbb{R}^d : |x|^2 \leqslant 1\}$$

i *sferze*

$$S^{d-1} = \{x \in \mathbb{R}^d : |x|^2 = 1\}.$$

Oczywiście, $|x|$ oznacza normę euklidesową, $|x| = (x_1^2 + \cdots + x_d^2)^{1/2} = (x^\top x)^{1/2}$. Rozkład $\mathrm{U}(B^d)$ ma po prostu stałą gęstość względem d-wymiarowej miary Lebesgue'a na kuli. Rozkład $\mathrm{U}(S^{d-1})$

ma stałą gęstość względem $(d-1)$-wymiarowej miary „powierzchniowej" na sferze. Oba te rozkłady są niezmiennicze względem *obrotów* (liniowych przekształceń ortogonalnych) $\mathbb{R}^d$. Takie rozkłady nazywamy *sferycznie symetrycznymi* lub krócej: **sferycznymi**. Zauważmy, że zmienną losową o rozkładzie $\mathrm{U}(S^{d-1})$ możemy interpretować jako losowo wybrany kierunek w przestrzeni $d-1$-wymiarowej. Algorytmy „poszukiwań losowych" często wymagają generowania takich losowych kierunków.

Rozkłady jednostajne na kuli i sferze są blisko ze sobą związane.

— Jeśli $V = (V_1, \ldots, V_d) \sim \mathrm{U}(B^d)$ i $R = |V|$ to

$$Y = \frac{V}{R} = \left(\frac{V_1}{R}, \ldots, \frac{V_d}{R}\right) \sim \mathrm{U}(S^{d-1}).$$

Latwo też zauważyć, że R jest zmienną losową o rozkładzie $\mathbb{P}(R \leqslant r) = r^d$ niezależną od Y.

— Jeśli $Y \sim \mathrm{U}(S^{d-1})$ i R jest niezależną zmienną losową o rozkładzie $\mathbb{P}(R \leqslant r) = r^d$ to $V = RY = (RY_1, \ldots, RV_d) \sim \mathrm{U}(B^d)$.

Zmienną R łatwo wygenerować metodą odwracania dystrybuanty.

Najprostszy wydaje się algorytm eliminacji:

Listing.

repeat
  Gen $V_1, \ldots, V_d \sim \mathrm{U}(0,1)$
until $R^2 = V_1^2 + \ldots + V_d^2 \leqslant 1$

Na wyjściu otrzymujemy, zgodnie z żądaniem

$$V = (V_1, \ldots, V_d) \sim \mathrm{U}(B^d).$$

W istocie, dokładnie ta metoda, dla $d = 2$, była częścią algorytmu biegunowego Marsaglii. Problem w tym, że w wyższych wymiarach efektywnosc eliminacji gwałtownie maleje. Prawdopodobieństwo akceptacji jest równe stosunkowi „objętości" kuli B^d do kostki $[-1,1]^d$. Ze znanego wzoru na objętość kuli d-wymiarowej wynika, że

$$\frac{|B^d|}{2^d} = \frac{2\pi^{d/2}}{d\Gamma(d/2)} \cdot \frac{1}{2^d} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \longrightarrow_{d\to\infty} 0.$$

Zbieżność do zera jest bardzo szybka. Dla dużego d kula jest znikomą częścią opisanej na niej kostki.

Inna metoda, którą z powodzeniem stosuje się dla $d = 2$ jest związana ze współrzędnymi biegunowymi:

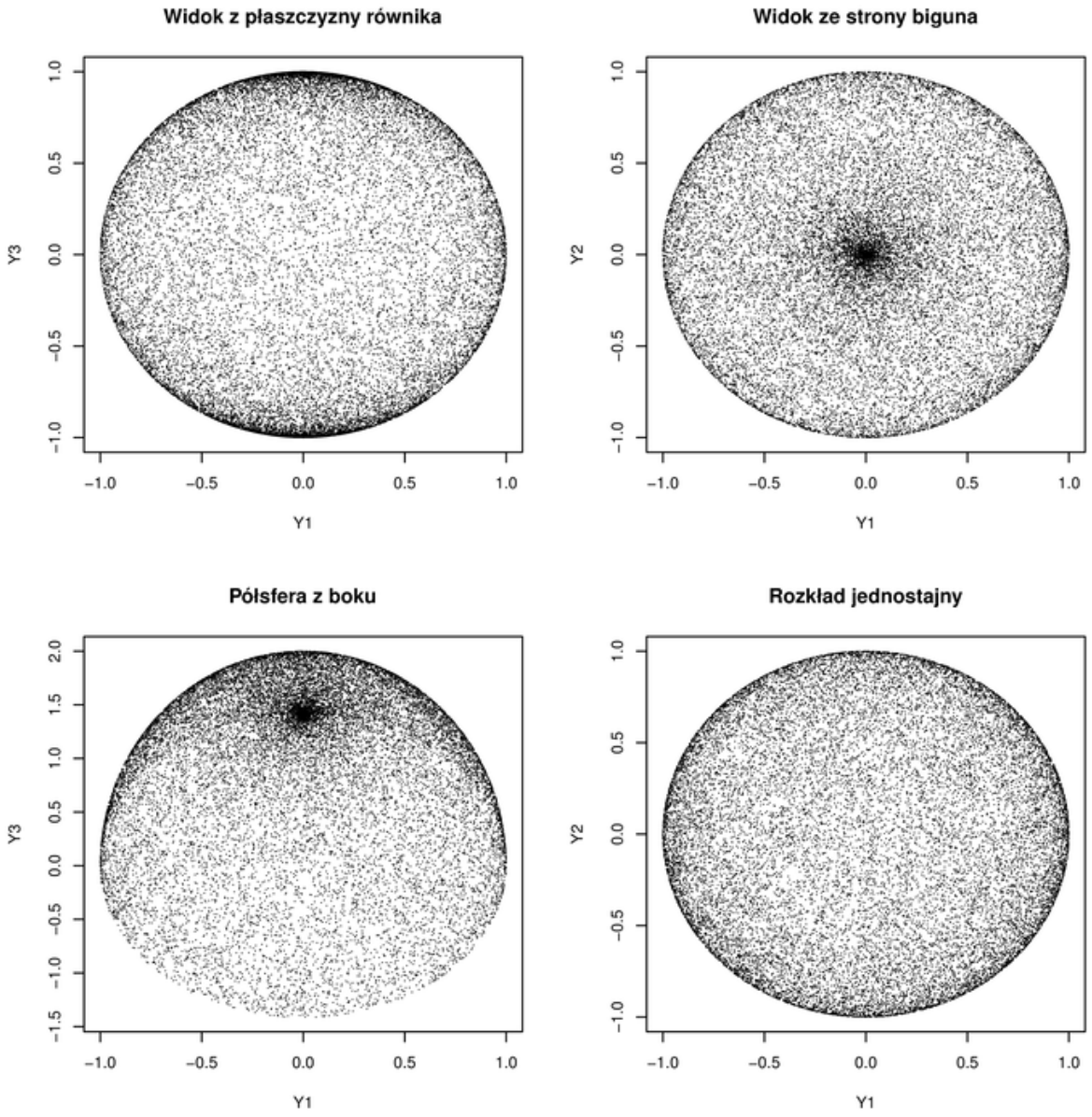
Listing.

Gen $\Phi \sim \mathrm{U}(0, 2\pi)$;
$Y_1 := \cos\Phi$; $Y_2 := \sin\Phi$;
Gen U; $R := \sqrt{U}$;
$V_1 := Y_1 \cdot R$; $V_2 := Y_2 \cdot R$;

Na wyjściu $(Y_1, Y_2) \sim S^1$ i $(V_1, V_2) \sim B^2$. Jest to część algorytmu Boxa-Müllera. Uogólnienie na przypadek $d > 2$ nie jest jednak ani proste, ani efektywne. Mechaniczne zastąpienie, współrzędnych biegunowych przez współrzędne sferyczne (dla, powiedzmy $d = 3$) prowadzi do *niepoprawnych wyników*. Popatrzmy na punkty produkowane przez następujący algorytm:

Listing.

Gen $\Phi \sim \mathrm{U}(0, 2\pi)$; $\Psi \sim \mathrm{U}\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$
$Y_1 := \cos\Phi\cos\Psi$; $Y_2 := \sin\Phi\cos\Psi$; $Y_3 = \sin\Psi$



Rysunek 5.2. Niepoprawne i poprawne generowanie z rozkładu jednostajnego na sferze.

Efekt działania algorytmu „współrzędnych sferycznych" jest widoczny na trzech początkowych obrazkach na Rysunku 5.2. Górny lewy rysunek przedstawia punkty widoczne „z płaszczyzny równika", czyli (Y_1, Y_3). Górny prawy – te same punkty widziane „znad bieguna", czyli (Y_1, Y_2). Dolny lewy – górną półsferę widzianą „skośnie", czyli (Y_2, Y_3^a), gdzie $Y_3^a = \sqrt{2}Y_3 + \sqrt{2}Y_1$. Dla porównania, na ostatnim rysunku po prawej stronie u dołu – punkty wylosowane rzeczywiście z rozkładu $\mathrm{U}(S^2)$ przy pomocy poprawnego algorytmu podanego niżej.

Listing.

Gen $Z_1, \ldots, Z_d \sim \mathrm{N}(0,1)$;
$R := (Z_1^2 + \cdots + Z_d^2)^{1/2}$;
$Y_1 := Z_1/R, \ldots, Y_d := Z_d/R$;
Gen U; $\quad R := U^{1/d}$;

$V_1 := Y_1 \cdot R, \ldots, V_d := Y_d \cdot R;$

Na wyjściu $Y \sim \mathrm{U}(S^{d-1})$ i $V \sim \mathrm{U}(B^d)$. Jak widać, algorytm polega na normowaniu punktów wylosowanych ze sferycznie symetrycznego rozkładu normalnego. Jest prosty, efektywny i godny polecenia.

Przykład 5.3 (Wielowymiarowe rozkłady Studenta)**.** Niech $Z = (Z_1, \ldots, Z_d)^\top$ będzie wektorem losowym o rozkładzie $\mathrm{N}(0, I)$, zaś R^2 – niezależną zmienną losową o rozkładzie $\chi^2(n)$. Wektor

$$(Y_1, \ldots, Y_d)^\top = \frac{(Z_1, \ldots, Z_d)^\top}{\sqrt{R^2/n}}$$

ma, z definicji, *Sferyczny rozkład t-Studenta z n stopniami swobody.* Gęstość tego rozkładu (z dokładnością do stałej normującej) jest równa

$$f(y) = f(y_1, \ldots, y_d) \propto \left[1 + \frac{1}{n}(\sum y_i^2)\right]^{-(n+d)/2} = \left[1 + \frac{1}{n}|y|^2\right]^{-(n+d)/2}.$$

W przypadku jednowymiarowym, a więc przyjmując $d = 1$, otrzymujemy dobrze znane rozklady t-Studenta z n stopniami swobody o gęstości

$$f(y) \propto \frac{1}{(1 + y^2/n)^{(n+1)/2}}$$

W szczególnym przypadku, biorąc za liczbę stopni swobody $n = 1$, otrzymujemy rozkłady Cauchy'ego. Na przykład, dwuwymiarowy rozkład Cauchy'ego ma taką gęstość:

$$f(y_1, y_2) \propto \frac{1}{(1 + y_1^2 + y_2^2)^{3/2}}.$$

Użytecznym uogólnieniem rozkładów sferycznych są rozkłady eliptyczne. Są one określone w następujący sposób. Niech Σ będzie macierzą symetryczną i nieosbliwą. Nazwijmy *uogólnionym obrotem* przekształcenie liniowe, które zachowuje uogólnioną normę $|x|_{\Sigma^{-1}} = (x^\top \Sigma^{-1} x)^{1/2}$. Rozkład jest z definicji *eliptycznie konturowany* lub krócej **eliptyczny**, gdy jest niezmienniczy względem uogólnionych obrotów (dla ustalonej macierzy Σ).

5.2.2. Rozklady Dirichleta

Definicja 5.1. Mówimy, że n-wymiarowa zmienna losowa X ma rozkład Dirichleta,

$$X = (X_1, \ldots, X_n) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$$

jesli $X_1 + \cdots + X_n = 1$ i zmienne $X_1, \ldots, X_{n-1}$ mają gęstość

$$f(x_1, \ldots, x_{n-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_n)} x_1^{\alpha_1 - 1} \ldots x_{n-1}^{\alpha_{n-1} - 1} (1 - x_1 - \ldots x_{n-1})^{\alpha_n - 1}.$$

Parametry $\alpha_1, \ldots, \alpha_n$ mogą być dowolnymi liczbami dodatnimi.

Uwaga 5.1. Rozkłady Dirichleta dla $d = 2$ są to w istocie rozkłady beta:

$$X_1 \sim \mathrm{Beta}(\alpha_1, \alpha_2) \text{ wtedy i tylko wtedy, gdy } (X_1, X_2) \sim \mathrm{Dir}(\alpha_1, \alpha_2)$$

Wniosek 5.1. *Jeśli $U_1, \ldots, U_{n-1}$ są niezależnymi zmiennymi o jednakowym rozkładzie jednostajnym* $\mathrm{U}(0,1)$ *i*

$$U_{1:n} < \ldots < U_{n-1:n-1}$$

oznaczają statystyki pozycyjne, to spacje

$$X_i = U_{i:n} - U_{i-1:n}, \quad X_n = 1 - U_{n-1:n}$$

maja rozklad $\mathrm{Dir}(1, \ldots, 1)$.

Twierdzenie 5.2. *Jeśli $Y_1, \ldots, Y_n$ są niezależnymi zmiennymi losowymi o rozkładach gamma,*

$$Y_i \sim \mathrm{Gamma}(\alpha_i)$$

i $S = Y_1 + \cdots + Y_n$ to

$$(X_1, \ldots, X_n) = \left(\frac{Y_1}{S}, \ldots, \frac{Y_n}{S}\right) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n).$$

Wektor losowy X jest niezależny od S.

Dowód. Obliczymy łączną gęstość zmiennych losowych $S, X_1, \ldots, X_{n-1}$. Ze wzoru na przekształcenie gęstości wynika, że

$$\begin{aligned} &f_{S,X_1,\ldots,X_{n-1}}(s, x_1, \ldots, x_{n-1}) = f_{Y_1,\ldots,Y_n}(x_1 s, \ldots, x_n s) \\ &\propto (x_1 s)^{\alpha_1 - 1} \mathrm{e}^{-x_1 s} \ldots (x_n s)^{\alpha_n - 1} \mathrm{e}^{-x_n s} \left| \frac{\partial(y_1, \ldots, y_n)}{\partial(s, x_1, \ldots x_{n-1})} \right| \\ &\propto x_1^{\alpha_1 - 1} \ldots x_n^{\alpha_n - 1} s^{\alpha_1 + \cdots + \alpha_n - 1} \mathrm{e}^{-s}, \end{aligned}$$

ponieważ jakobian przekształcenia odwrotnego jest równy s^{n-1}. Wystarczy teraz zauważyć, że

$$x_1^{\alpha_1 - 1} \cdots x_n^{\alpha_n - 1} = \mathrm{Dir},$$

$$s^{\alpha_1 + \cdots + \alpha_n - 1} \mathrm{e}^{-s} = \mathrm{Gamma}.$$

□

Wniosek 5.2. *Dla niezależnych zmiennych losowych o jednakowym rozkładzie wykładniczym,*

$$Y_1, \ldots, Y_n \sim \mathrm{Ex}(1),$$

jeśli $S = Y_1 + \cdots + Y_n$ to

$$(X_1, \ldots, X_n) = \left(\frac{Y_1}{S}, \ldots, \frac{Y_n}{S}\right) \sim \mathrm{Dir}(1, \ldots, 1)$$

Z 5.1 i 5.2 wynika, że następujące dwa algorytmy:

Listing.

Gen $U_1, \ldots, U_{n-1}$;
Sort $(U_1, \ldots, U_{n-1})$; $U_0 = 0$; $U_n = 1$;
$X_i := U_i - U_{i-1}$

oraz

Listing.

```
Gen Y_1, ..., Y_n ~ Ex(1)
S := Y_1 + ... + Y_n
X_i := Y_i / S
```

dają te same wyniki.

Wiele ciekawych własności rozkładów Dirichleta wynika niemal natychmiast z 5.2 (choć nie tak łatwo wyprowadzić je posługując się gęstością 5.2). Mam na myśli przede wszystkim zasadniczą własność „grupowania zmiennych".

Wniosek 5.3. *Rozważmy rozbicie zbioru indeksów na sumę rozłącznych podzbiorów:*

$$\{1,\dots,n\} = \bigcup_{j=1}^{k} I_j.$$

Jeżeli $(X_1,\dots,X_n) \sim \mathrm{Dir}(\alpha_1,\dots,\alpha_n)$ *i rozważymy „zgrupowane zmienne"*

$$S_j = \sum_{i\in I_j} X_i,$$

to wektor tych zmiennych ma też rozkład Dirichleta,

$$(S_1,\dots,S_k) \sim \mathrm{Dir}(\beta_1,\dots,\beta_k),$$

gdzie

$$\beta_j = \sum_{i\in I_j} \alpha_i.$$

Co więcej, każdy z wektorów $(X_i/S_j)_{i\in I_j}$ *ma rozkład Dirichleta* $\mathrm{Dir}(\alpha_i)_{i\in I_j}$ *i wszystkie te wektory są niezależne od* $(S_1,\dots,S_k)$.

Przykład 5.4. Wniosek 5.3 razem z 5.1 pozwala szybko generować wybrane statystyki pozycyjne. Na przykład łączny rozkład dwóch statystyk pozycyjnych z rozkładu jednostajnego jest wyznaczony przez rozkład trzech „zgrupowanych spacji":

$$(U_{k:n-1},\ U_{l:n-1} - U_{k:n-1},\ 1 - U_{l:n-1}) \sim \mathrm{Dir}(k, l-k, n-l)$$

Przykład 5.5 (Rozkład dwumianowy)**.** Aby wygenerować zmienną o rozkładzie dwumianowym $\mathrm{Bin}(n,p)$ wystarczy rozpoznać między którymi statystykami pozycyjnymi z rozkładu $\mathrm{U}(0,1)$ leży liczba p. Nie musimy w tym celu generować *wszystkich* staystyk pozycyjnych, możemy wybierać „najbardziej prawdopodobne". W połączeniu 5.4 daje to następujący algorytm.

Listing.

```
k := n; θ := p; X := 0;
repeat
  i := ⌊1 + kθ⌋;
  Gen V ~ Beta(i, k + 1 − i);
  if θ < V then
    begin θ := θ/V; k := i − 1 end
                  else
    begin X := X + i; θ = (θ − V)/(1 − V); k := k − i end

until k = 0
```

Metoda generowania zmiennych o rozkładzie Dirichleta opiera sie na następującym fakcie, który jest w istocie szczególnym przypadkiem „reguły grupowania” 5.3.

Wniosek 5.4. *Jeśli* $(X_1, \ldots, X_n) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$ *i, dla* $k = 1, \ldots, n$*, określimy* $S_k = X_1 + \cdots + X_k$ *to zmienne*

$$Y_1 = \frac{S_1}{S_2}, Y_2 = \frac{S_2}{S_3}, \ldots, Y_n = \frac{S_{n-1}}{S_n}$$

są niezależne i

$$Y_k \sim \mathrm{Beta}(\alpha_1 + \cdots + \alpha_k, \alpha_{k+1}).$$

Odwrotnie, jeśli zmienne $Y_1, \ldots, Y_n$ *są niezależne i każda z nich ma rozkład beta, to wektor* $(X_1, \ldots, X_n)$ *ma rozkład Dirichleta.*

Oczywiście, jeśli wygenerujemy niezależne zmienne $Y_1, \ldots, Y_n$ o rozkładach beta, to zmienne $X_1, \ldots, X_n$ łatwo „odzyskać” przy pomocy wzorów:

$$X_n = 1 - Y_{n-1}$$

$$Z_{n-1} = (1 - Y_{n-2})Y_{n-1}$$

$$\ldots$$

$$Z_2 = (1 - Y_1)Y_2 \cdots Y_{n-1}$$

$$Z_1 = Y_1 Y_2 \cdots Y_{n-1}$$

Powyższe równania określają algorytm generowania zmiennych o rozkładzie $\mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$.

6. Symulowanie procesów stochastycznych I.

Pełny tytuł tego rozdziału powinien brzmieć „Symulacje Niektórych Procesów Stochastycznych, Bardzo Subiektywnie Wybranych Spośród Mnóstwa Innych". Nie będę szczegółowo tłumaczył, skąd pochodzi mój subiektywny wybór. Zrezygnowałem z próby przedstawienia procesów z czasem ciągłym i równocześnie ciągłą przestrzenią stanów, bo to temat oddzielny i obszerny.

6.1. Stacjonarne procesy Gaussowskie

Ograniczymy się do dwóch klas procesów, często używanych do modelowania różnych zjawisk. Będą to procesy z czasem dyskretnym i przestrzenią stanów $\mathbb{R}$, to znaczy ciągi (zależnych) zmiennych losowych $X_0, X_1, \ldots, X_n, \ldots$ o wartościach rzeczywistych. Niech $\ldots, W_{-1}, W_0, W_1, \ldots, W_n, \ldots$ będzie ciągiem *niezależnych zmiennych losowych* o jednakowym rozkładzie $\mathrm{N}(0, v^2)$ (wygodnie posłużyć się tutaj ciągiem indeksowanym wszystkimi liczbami całkowitymi).

Definicja 6.1. Proces **ruchomych średnich** rzędu q, w skrócie $\mathrm{MA}(q)$ jest określony równaniem

$$X_n = \beta_1 W_{n-1} + \cdots + \beta_q W_{n-q}, \qquad (n = 0, 1, \ldots),$$

gdzie $\beta_1, \ldots, \beta_q$ jest ustalonym ciągiem współczynników.

Sposób generowania takiego procesu jest oczywisty i wynika wprost z definicji. Co więcej widać, że proces $\mathrm{MA}(q)$ jest *stacjonarny*, to znaczy łączny rozkład prawdopodobieństwa zmiennych $X_0, X_1, \ldots, X_n$ jest taki sam jak zmiennych $X_k, X_{k+1}, \ldots, X_{k+n-1}$, dla dowolnych n i k. Intuicyjnie, proces nie zmienia się po „przesunięciu czasu" o k jednostek.

Definicja 6.2. Proces **autoregresji** rzędu p, w skrócie $\mathrm{AR}(p)$ jest określony równaniem rekurencyjnym

$$X_n = \alpha_1 X_{n-1} + \cdots + \alpha_p X_{n-p} + W_n, \qquad (n = p, p+1, \ldots),$$

gdzie $\alpha_1, \ldots, \alpha_p$ jest ustalonym ciągiem współczynników.

Procesy autoregresji wydają się bardzo odpowiednie do modelowania „szeregów czasowych": stan układu w chwili n zależy od stanów przeszłych i dotatkowo jeszcze od przypadku. Procesy $\mathrm{AR}(1)$, w szczególności są łańcuchami Markowa. Sposób generowania procesów $\mathrm{AR}(p)$ jest też bezpośrednio widoczny z definicji. Pojawia się jednak pewien problem. Jak znaleźć $X_0, \ldots, X_{p-1}$ na początku algorytmu w taki sposób, żeby proces był stacjonarny? Jest to o tyle istotne, że rzeczywiste procesy (na przykład czeregi czasowe w zastosowaniach ekonomicznych) specjaliści uznają za stacjonarne, przynajmniej w przybliżeniu.

Rozważmy dla wygody oznaczeń podwójnie nieskończony proces

$$\ldots, X_{-1}, X_0, X_1, \ldots$$

spełniający równanie autoregresji rzędu p. Załóżmy, że ten proces jest stacjonarny i wektor $X_0, \ldots, X_{p-1}$ rozkład normalny, $\mathrm{N}(0, \Sigma)$. Stacjonarność implikuje, że elementy macierzy Σ muszą być postaci $\mathrm{Cov}(X_i, X_j) = \sigma^2 \rho_{i-j}$. Mamy przy tym $\rho_{-k} = \rho_k$, co może być traktowane jako

wygodna konwencja (po to właśnie „rozszerzamy" proces w obie strony). Zastosowanie równania definiującego autoregresję prowadzi do wniosku, że

$$\begin{aligned}\sigma^2\rho_k &= \mathrm{Cov}(X_0, X_k) = \mathrm{Cov}(X_0, \alpha_1 X_{k-1} + \cdots + \alpha_p X_{k-p} + W_k)\\ &= \sigma^2\alpha_1\rho_{k-1} + \sigma^2\alpha_2\rho_{k-2} + \cdots + \sigma^2\alpha_p\rho_{k-p}.\end{aligned}$$

Podobnie,

$$\begin{aligned}\sigma^2 &= \mathrm{Var}(X_0) = \mathrm{Cov}(X_0, \alpha_1 X_{-1} + \cdots + \alpha_p X_{-p} + W_0)\\ &= \sigma^2\alpha_1\rho_1 + \sigma^2\alpha_2\rho_2 + \cdots + \sigma^2\alpha_p\rho_p + v^2,\end{aligned}$$

gdzie $v^2 = \mathrm{Var}(W_0)$. Otrzymujemy następujący układ równań na współczynniki autokorelacji ρ_k:

$$\begin{cases}\rho_1 = \alpha_1 + \alpha_2\rho_1 + \alpha_3\rho_2 + \ldots + \alpha_p\rho_{p-1},\\ \rho_2 = \alpha_1\rho_1 + \alpha_2 + \alpha_3\rho_1 + \ldots + \alpha_p\rho_{p-2},\\ \ldots,\\ \rho_p = \alpha_1\rho_{p-1} + \alpha_2\rho_{p-2} + \alpha_3\rho_2 + \ldots + \alpha_p,\end{cases} \tag{6.1}$$

Można pokazać, że ten ten układ ma rozwiązanie, jeśli wielomian charakterystyczny $A(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p$ nie ma zer w kole $\{|z| \leqslant 1\}$. Ponadto mamy równanie na wariancję stacjonarną:

$$\sigma^2 = \frac{v^2}{1 - \rho_1\alpha_1 - \ldots - \alpha_p\rho_p} \tag{6.2}$$

Metoda generowania *stacjonarnego* procesu $\mathrm{AR}(p)$, $X_0, X_1, \ldots, X_p, \ldots$ jest następująca. Znajdujemy rozwiązanie układu równań (6.1), wariancję obliczamy ze wzoru (6.2) i tworzymy macierz $\Sigma = (\sigma^2\rho_{i-j})_{i,j=0,\ldots,p-1}$. Generujemy wektor losowy $(X_0, X_1, \ldots, X_{p-1}) \sim \mathrm{N}(0, \Sigma)$ i dalej generujemy rekurencyjnie $X_p, X_{p+1}, \ldots$ używając równania autoregresji. Aby się przekonać, że tak generowany proces jest stacjonarny, wystarczy sprawdzić że identyczne są rozkłady wektorów $(X_0, X_1, \ldots, X_{p-1})$ i $(X_1, X_2, \ldots, X_p)$. Mamy

$$\begin{pmatrix}X_1\\ .\\ .\\ .\\ X_p\end{pmatrix} = \begin{pmatrix}0 & 1 & 0 & \ldots & 0\\ 0 & 0 & 1 & \ldots & 0\\ 0 & \ddots & \ddots & \ddots & 0\\ 0 & \ldots & \ldots & 0 & 1\\ \alpha_p & \ldots & \ldots & \ldots & \alpha_1\end{pmatrix}\begin{pmatrix}X_0\\ .\\ .\\ .\\ X_{p-1}\end{pmatrix} + \begin{pmatrix}0\\ .\\ .\\ .\\ W_p\end{pmatrix}.$$

Niech R oznacza „dużą macierz" w tym wzorze . Z własności wielowymiarowych rozkładów normalnych wynika, że wystarczy sprawdzić równość

$$\Sigma = R\Sigma R^T + \begin{pmatrix}0 & 0 & \ddots & 0\\ 0 & \ddots & \ddots & 0\\ 0 & \ddots & \ddots & 0\\ 0 & 0 & \ldots & v^2\end{pmatrix}.$$

Macierz Σ została tak wybrana, że ta równość jest spełniona.

6.2. Procesy Poissona

6.2.1. Jednorodny proces Poissona na półprostej

Definicja 6.3. Rozważmy niezależne zmienne losowe $W_1, \ldots, W_k, \ldots$ o jednakowym rozkładzie wykładniczym, $X_i \sim \mathrm{Ex}(\lambda)$ i utwórzmy kolejne sumy

$$T_0 = 0,\ T_1 = W_1,\ T_2 = W_1 + W_2,\ \ldots,\ T_k = W_1 + \cdots + W_k,\ \ldots$$

Niech, dla $t \geqslant 0$,

$$N(t) = \max\{k : T_k \leqslant t\}.$$

Rodzinę zmiennych losowych $N(t)$ nazywamy *procesem Poissona*.

Proces Poissona dobrze jest wyobrażać sobie jako *losowy zbiór punktów* na półprostej: $\{T_1, T_2, \ldots, T_k, \ldots\}$. Zmienna $N(t)$ oznacza liczbę punktów, które „wpadły" w odcinek $]0, t]$. Wygodnie będzie używać symbolu

$$N(s,t) = N(t) - N(s)$$

dla oznaczenia liczby punktów, które „wpadły" w odcinek $]s, t]$.

Stwierdzenie 6.1. *Jeśli $N(t)$ jest procesem Poissona, to*

$$\mathbb{P}(N(t) = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}.$$

Dowód. Zauważmy, że

$$T_k \sim \mathrm{Gamma}(k, \lambda).$$

Wobec tego ze wzoru na prawdopodobieństwo całkowite wynika, że

$$\begin{aligned}
\mathbb{P}(N(t) = k) &= \mathbb{P}(T_k \leqslant t, T_{k+1} > t) \\
&= \int_0^t \mathbb{P}(T_{k+1} > t | T_k = s) f_{T_k}(s)\ \mathrm{d}s \\
&= \int_0^t \mathbb{P}(W_{k+1} > t - s | T_k = s) f_{T_k}(s)\ \mathrm{d}s \\
&= \int_0^t \mathrm{e}^{-\lambda(t-s)}\frac{\lambda^k}{\Gamma(k)} s^{k-1}\mathrm{e}^{-\lambda s}\mathrm{d}s \\
&= \mathrm{e}^{-\lambda t}\frac{\lambda^k}{\Gamma(k)}\int_0^t s^{k-1}\mathrm{d}s = \mathrm{e}^{-\lambda t}\frac{\lambda^k}{(k-1)!}\frac{t^k}{k} = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}.
\end{aligned}$$

□

Oczywiście $\mathbb{E}N(t) = \lambda t$. Liczba

$$\lambda = \frac{1}{\mathbb{E}W_i} = \frac{\mathbb{E}N(t)}{t}$$

jest nazywana *intensywnością procesu*.

Stwierdzenie 6.2. *Jeśli*

$$0 < t_1 < t_2 < \cdots < t_i < \cdots,$$

to zmienne losowe $N(t_1), N(t_1, t_2), \ldots$ są niezależne i każda z nich ma rozkład Poissona:

$$N(t_{i-1}, t_i) \sim \mathrm{Poiss}(\lambda(t_i - t_{i-1})).$$

Dowód. Pokażemy, że warunkowo, dla $N(t_1) = k$, ciąg zmiennych losowych

$$T_{k+1} - t_1, W_{k+2}, W_{k+3} \ldots, \quad \text{jest iid} \sim \mathrm{Ex}(\lambda).$$

Wynika to z własności braku pamięci rozkładu wykładniczego. W istocie, dla ustalonych $N(t_1) = k$ i $S_k = s$ mamy

$$\begin{aligned} &\mathbb{P}(T_{k+1} - t_1 > t | N(t_1) = k, T_k = s) \\ &= \mathbb{P}(T_{k+1} - t_1 > t | T_k = s, T_{k+1} > t_1) \\ &= \mathbb{P}(W_{k+1} > t_1 + t - s | W_{k+1} > t_1 - s) \\ &= \mathbb{P}(W_{k+1} > t) = \mathrm{e}^{-\lambda t}. \end{aligned}$$

Fakt, że zmienne $W_{k+2}, W_{k+3} \ldots$ są niezależne od zdarzenia $N(t_1) = k$ jest oczywisty. Pokazaliśmy w ten sposób, że losowy zbiór punktów $\{T_{k+1} - t_1, T_{k+2} - t_1, \ldots\}$ ma (warunkowo, dla $N(t_1) = k$) taki sam rozkład prawdopodobieństwa, jak $\{T_1, T_2, \ldots\}$. Proces Poissona *obserwowany od momentu t_1 jest kopią wyjściowego procesu.* Wynika stąd w szczególności, że zmienna losowa $N(t_2, t_1)$ jest niezależna od $N(t_1)$ i $N(t_2, t_1) \sim \mathrm{Poiss}(\lambda(t_2 - t_1))$. Dalsza część dowodu przebiega analogicznie i ją pominiemy. □

Metoda generowania procesu Poisson oparta na Definicji 6.3 jest raczej oczywista. Zauważmy jednak, że nie jest to *jedyna* metoda. Inny sposób generowania (i inny sposób patrzenia na proces Poissona) jest związany z następującym faktem.

Stwierdzenie 6.3. *Warunkowo, dla $N(t) = n$, ciąg zmiennych losowych*

$$T_1, \ldots, T_n$$

ma rozkład taki sam, jak ciąg statystyk pozycyjnych

$$U_{1:n}, \ldots, U_{n:n}$$

z rozkładu $U(0, t)$.

Dowód. Z Wniosku 5.1 wynika, że warunkowo, dla $T_n = s$,

$$\left(\frac{W_1}{s}, \ldots, \frac{W_n}{s}\right) \sim \mathrm{Dir}(\underbrace{1, \ldots, 1}_{n}).$$

Innymi słowy wektor losowy $(T_1, \ldots, T_{n-1})$ ma taki rozkład, jak $n-1$ statystyk pozycyjnych z $U(0, s)$.

Obliczmy warunkową gęstość zmiennej losowej T_n, jeśli $N(t) = n$:

$$\begin{aligned} f_{T_n}(s | N(t) = n) &= \frac{\mathbb{P}(N(t) = n | T_n = s) f_{T_n}(s)}{\mathbb{P}(N(t) = n)} \\ &= \frac{\mathbb{P}(W_{n+1} > t - s) f_{T_n}(s)}{\mathbb{P}(N = n)} \\ &= \frac{\mathrm{e}^{-\lambda(t-s)}(\lambda^n/(n-1)!) s^{n-1} \mathrm{e}^{-\lambda s}}{\mathrm{e}^{-\lambda t}(\lambda t)^n / n!} \\ &= \frac{n s^{n-1}}{t^n}. \end{aligned}$$

A zatem warunkowo, dla $N(t) = n$, zmienna losowa T_n/t ma rozklad Beta$(n, 1)$ i w konsekwencji (przypomnijmy sobie algorytm generowania zmiennych o rozkładzie Dirichleta)

$$\left(\frac{W_1}{t}, \ldots, \frac{W_n}{t}, \frac{t - T_n}{t}\right) \sim \mathrm{Dir}(\underbrace{1, \ldots, 1}_{n+1}).$$

Innymi słowy wektor losowy $(T_1, \ldots, T_{n-1}, T_n)$ ma taki rozkład, jak n statystyk pozycyjnych z $U(0, t)$. □

Wynika stąd następujący sposób generowania procesu Poissona na przedziale $[0, t]$.

Listing.

Gen $N \sim \mathrm{Poiss}(\lambda t)$
for $i = 1$ **to** N **do** Gen $U_i \sim \mathrm{U}(0, t)$;
Sort $(U_1, \ldots, U_N)$
$(T_1, \ldots, T_N) := (U_{1:N}, \ldots, U_{N:N})$

Co ważniejsze, Stwierdzenia 6.1, 6.2 i 6.3 wskazują, jak powinny wyglądać uogólnienia procesu Poissona i jak generować takie ogólniejsze procesy. Zanim tym się zajmiemy, zróbmy dygresję i podajmy twierdzenie charakteryzujące „zwykły" proces Poissona. Nie jest ono bezpośrednio używane w symulacjach, ale wprowadza pewien ważny sposób określania procesów, który ułatwia zrozumienie łańcuchów Markowa z czasem ciągłym i jest bardzo użyteczny w probabilistycznym modelowaniu zjawisk.

Twierdzenie 6.1. *Załóżmy, że $(N(t) : t \geqslant 0)$ jest procesem o wartościach w $\{0, 1, 2, \ldots\}$, stacjonarnych i niezależnych przyrostach (to znaczy $N(t) - N(s)$ jest niezależne od $(N(u), u \leqslant s)$ i ma rozkład zależny tylko od $t - s$ dla dowolnych $0 < s < t$) oraz, że trajektorie $N(t)$ są prawostronnie ciągłymi funkcjami mającymi lewostronne granice (prawie na pewno). Jeżeli $N(0) = 0$ i spełnione są następujące warunki:*

$$(i) \quad \lim_{t \to 0} \frac{\mathbb{P}(N(t) = 1)}{t} = \lambda,$$

$$(ii) \quad \lim_{t \to 0} \frac{\mathbb{P}(N(t) \geqslant 2)}{t} = 0,$$

to $N(\cdot)$ jest jednorodnym procesem Poissona z intensywnością λ

Bardzo prosto można zauważyć, że proces Poissona $(N(t) : t \geqslant 0)$ o intensywności λ ma własności wymienione w Twierdzeniu 6.1. Ciekawe jest, że te własności *w pełni charakteryzują* proces Poissona.

Dowód Tw. 6.1 – szkic. Pokażemy tylko, że

$$p_n(t) := \mathbb{P}(N(t) = n) = \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!}.$$

Najpierw zajmiemy się funkcją $p_0(t) = \mathbb{P}(N(t) = 0)$. Z niezależności i jednorodności przyrostów wynika tożsamość

$$p_0(t + h) = p_0(h) p_0(t).$$

Stąd

$$\frac{p_0(t+h)-p_0(t)}{h} = \frac{p_0(h)-1}{h}p_0(t) = \Big[-\frac{p_1(h)}{h} - \frac{\sum_{i=2}^{\infty} p_i(h)}{h}\Big]p_0(h).$$

Przejdźmy do granicy z $h \to 0$ i skorzystajmy z własności (i) i (ii). Dostajemy proste równanie różniczkowe:

$$p_0'(t) = -\lambda p_0(t).$$

Rozwiązanie tego równania z warunkiem początkowym $p_0(0) = 1$ jest funkcja

$$p_0(t) = \mathrm{e}^{-\lambda t}.$$

Bardzo podobnie obliczamy kolejne funkcje p_n. Postępujemy rekurencyjnie: zakładamy, że znamy p_{n-1} i układamy równanie różniczkowe dla funkcji p_n. Podobnie jak poprzednio,

$$p_n(t+h) = p_n(t)p_0(h) + p_{n-1}(t)p_1(h) + \sum_{i=2}^{n} p_{n-1}(t)p_i(h),$$

a zatem

$$\frac{p_n(t+h)-p_n(t)}{h} = \frac{p_0(h)-1}{h}p_n(t) + \frac{p_1(h)}{h}p_{n-1}(t) + \frac{1}{h}\sum_{i=2}^{n} p_{n-i}(t)p_i(h).$$

Korzystając z własności (i) i (ii) otrzymujemy równanie

$$p_n'(t) = -\lambda p_n(t) + \lambda p_{n-1}(t).$$

To równanie można rozwiązać metodą uzmiennienia stałej: poszukujemy rozwiązania postaci $p_n(t) = c(t)\mathrm{e}^{-\lambda t}$. Zakładamy przy tym indukcyjnie, że $p_{n-1}(t) = (\lambda t)^{n-1}\mathrm{e}^{-\lambda t}/(n-1)!$ i mamy oczywisty warunek początkowy $p_n(0) = 0$. Stąd już łatwo dostać dowodzony wzór na p_n.

Na koniec zauważmy, że z postaci funkcji p_0 łatwo wywnioskować jaki ma rozkład zmienna $T_1 = \inf\{t : N(t) > 0\}$. Istotnie, $\mathbb{P}(T_1 > t) = \mathbb{P}(N(t) = 0) = p_0(t) = \mathrm{e}^{-\lambda t}$. □

Własności (i) i (ii), w połączeniu z jednorodnością przyrostów można przepisać w następującej sugestywnej formie:

$$\begin{aligned} &\mathbb{P}(N(t+h) = n+1 | N(t) = n) = \lambda h + o(h), \\ &\mathbb{P}(N(t+h) = n | N(t) = n) = 1 - \lambda h + o(h), \quad h \searrow 0. \end{aligned} \tag{6.3}$$

Jest to o tyle ważne, że w podobnym języku łatwo formułować założenia o ogólniejszych procesach, na przykład tak zwanych procesach urodzin i śmierci. Wrócimy do tego przy okazji omawiania łańcuchów Markowa.

6.2.2. Niejednorodne procesy Poissona w przestrzeni

Naturalne uogólnienia procesu Poissona polegają na tym, że rozważa się losowe zbiory punktów w przestrzeni o dowolnym wymiarze i dopuszcza się różną intensywność pojawiania się punktów w różnych rejonach przestrzeni. Niech $\mathcal{X}$ będzie przestrzenią polską. Czytelnik, który nie lubi abstrakcji może założyć, że $\mathcal{X} \subseteq \mathbb{R}^d$.

Musimy najpierw wprowadzić odpowiednie oznaczenia. Rozważmy ciąg wektorów losowych w $\mathcal{X}$:

$$X_1, \ldots, X_n, \ldots$$

(może to być ciag skonczony lub nie, liczba tych wektorów może być zmienną losową). Niech, dla $A \subseteq \mathcal{X}$,

$$N(A) = \#\{i : X_i \in A\}$$

oznacza liczbę wektorów, które „wpadły do zbioru A" (przy tym dopuszczamy wartość $N(A) = \infty$ i umawiamy się liczyć powtarzające się wektory tyle razy, ile razy występują w ciągu).

Niech teraz μ będzie miarą na (σ-ciele borelowskich podzbiorów) przestrzeni $\mathcal{X}$.

Definicja 6.4. $N(\cdot)$ jest procesem Poissona z miara intensywnosci μ, jesli

— dla parami rozłącznych zbiorów $A_1, \ldots, A_i \subseteq \mathcal{X}$, odpowiadające im zmienne losowe $N(A_1), \ldots, N(A_i)$ są niezależne;

— $\mathbb{P}(N(A) = n) = \mathrm{e}^{-\mu(A)} \dfrac{\mu(A)^n}{n!}$, dla każdego $A \subseteq \mathcal{X}$ takiego, że $\mu(A) < \infty$ i dla $n = 0, 1, \ldots$.

Z elementarnych własności rozkładu Poissona wynika następujący wniosek.

Wniosek 6.1. *Rozważymy rozbicie zbioru A o skończonej mierze intensywności, $\mu(A) < \infty$, na rozłączną sumę $A = A_1 \cup \cdots \cup A_k$. Wtedy*

$$\begin{aligned}\mathbb{P}\left(N(A_1) = n_1, \ldots, N(A_k) = n_k | N(A) = n\right) \\ = \frac{n!}{n_1! \cdots n_k!} \mu(A_1)^{n_1} \cdots \mu(A_k)^{n_k}, \qquad (n_1 + \cdots + n_k = n).\end{aligned}$$

Jeśli natomiast $\mu(A) = \infty$ to łatwo zauważyć, że $N(A) = \infty$ z prawdopodobieństwem 1.

Z Definicji 6.4 i Wniosku 6.1 natychmiast wynika następujący algorytm generowania procesu Poissona. Załóżmy, że interesuje nas „fragment" procesu w zbiorze $A \subseteq \mathcal{X}$ o skończonej mierze intensywności. W praktyce zawsze symulacje muszą się do takiego fragmentu ograniczać. Zauważmy, że *unormowana* miara $\mu(\cdot)/\mu(A)$ jest rozkładem prawdopodobieństwa (zmienna losowa X ma ten rozkład, jeśli $\mathbb{P}(X \in B) = \mu(B)/\mu(A)$, dla $B \subseteq A$).

Listing.

```
Gen N ~ Poiss(μ(A))
for i = 1 to N do Gen X_i ~ μ(·)/μ(A)
```

Należy rozumieć, że formalnie definiujemy $N(B) = \#\{i : X_i \in B\}$ dla $B \subseteq A$, w istocie jednak za realizację procesu uważamy zbiór punktów $\{X_1, \ldots, X_N\} \subset A$ („zapominamy" o uporządkowaniu punktów). Widać, że to jest proste uogólnienie analogicznego algorytmu dla „zwykłego" procesu Poissona, podanego wcześniej.

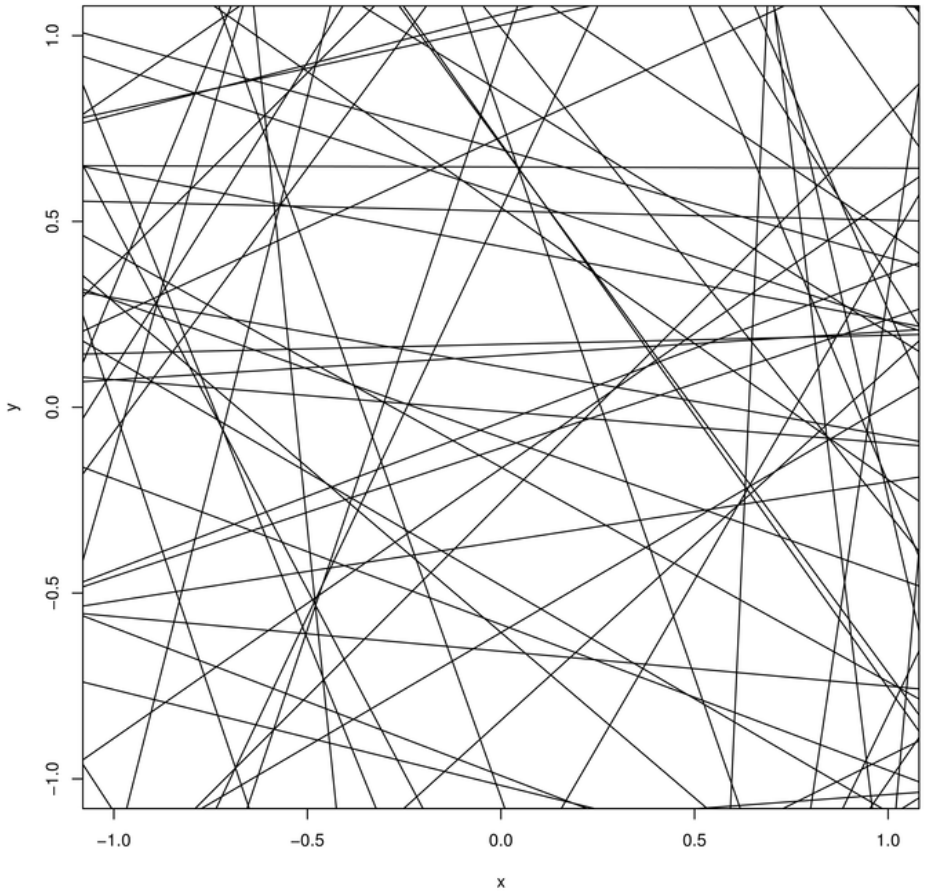
Można również rozbić zbiór A na rozłączną sumę $A = A_1 \cup \cdots \cup A_k$ i generować *niezależnie* fragmenty procesu w każdej części A_j.

Przykład 6.1. Jednorodny proces Poissona na kole $B^2 = \{x^2 + y^2 \leqslant 1\}$ można wygenerować w następujący sposób. Powiedzmy, że intensywność (punktów na jednostkę pola) jest λ, to znaczy $\mu(B) = \lambda|B|$, gdzie $|B|$ jest polem (miarą Lebesgue'a) zbioru B.

Najpierw generujemy $N \sim \mathrm{Poiss}(\lambda\pi)$, następnie punkty $(X_1, Y_1), \ldots, (X_N, Y_N)$ niezależnie z rozkładu $\mathrm{U}(B^2)$. To wszystko.

Ciekawszy jest następny przykład.

Przykład 6.2 (Jednorodny proces Poissona w przestrzeni prostych)**.** Prostą na płaszczyźnie można sparametryzować podając kąt $\varphi \in [0, 2\pi[$ jako tworzy prostopadła do prostej z osią poziomą oraz odległość $r \geqslant 0$ prostej od początku układu. Każda prosta jest więc opisana przez parę liczb (φ, r) czyli punkt przestrzeni $\mathcal{L} = [0, 2\pi[\times [0, \infty[$. Jeśli teraz wygenerujemy jednorodny proces Poissona na tej przestrzeni, to znaczy proces o intensywności $\mu(B) = \lambda|B|$, $B \subseteq \mathcal{L}$, to można się spodziewać zbioru „losowo położonych prostych". To widać na Rysunku 6.1. W istocie, wybór parametryzacji zapewnia, że rozkład prawdopodobieństwa procesu nie zależy od wyboru układu współrzędnych. Dowód, czy nawet precyzyjne sformułowanie tego stwierdzenia przekracza ramy tego skryptu. Intuicyjnie chodzi o to, że na obrazku „nie ma wyróżnionego



Rysunek 6.1. Proces Poissona w przestrzeni prostych.

kierunku ani wyróżnionego punktu". Nie można sensownie zdefiniować pojęcia „losowej prostej" ale każdy przyzna, że proces Poissona o którym mowa można uznać za uściślenie intuicyjnie rozumianego pojęcia „losowego *zbioru* prostych".

Ciekawe, że podstawowe metody generowania zmiennych losowych mają swoje odpowiedniki dla procesów Poissona. Rolę rozkładu prawdopodobieństwa przejnuje miara intensywności.

Przykład 6.3 (Odwracanie dystrybuanty)**.** Dla miary intensywności λ na przestrzeni jednowymiarowej można zdefiniować *dystrybuantę* tej miary. Dla uproszczenia rozważmy przestrzeń $\mathcal{X} = [0, \infty[$ i założymy, że każdy zbiór ograniczony ma miarę skończoną. Niech $\Lambda(t) = \lambda([0, t])$. Funkcję $\Lambda : [0, \infty[\to [0, \infty[$ nazwiemy dystrybuantą. Jest ona niemalejąca, prawostronnie ciągła, ale granica w nieskończoności $\Lambda(\infty) = \lim_{x\to\infty} \Lambda(x)$ może być dowolnym elementem z $[0, \infty]$. Dla procesu Poissona na $[0, \infty[$ wygodnie wrócić do prostszych oznaczeń, pisząc $N(t) = N([0, t])$ jak we wcześniej rozpatrywanym przypadku jednorodnym.

Niech $J(t)$ będzie jednorodnym procesem Poissona na $[0, \infty[$ z intensywnoscia równą 1. Wtedy

$$N(t) = J(\Lambda(t))$$

jest niejednorodnym procesem Poissona z miarą intensywności λ. W istocie, jeśli $0 < t_1 < t_2 < \cdots$ to $N(t_1), N(t_2) - N(t_1), \ldots$ są niezależne i maja rozkłady odpowiednio $\text{Poiss}(\Lambda(t_1)), \text{Poiss}(\Lambda(t_2) -$

$\Lambda(t_1)), \ldots$. Zauważmy, że jeśli $R_1 < R_2 < \cdots$ oznaczają punkty skoku procesu $J(\cdot)$ to $N(t) = \max\{k : R_k \leqslant \Lambda(t)\} = \max\{k : \Lambda^-(R_k) \leqslant t\}$, gdzie Λ^- jest uogólnioną funkcją odwrotną do dystrybuanty. Wobec tego punktami skoku procesu N są $T_k = \Lambda^-(R_k)$. Algorytm jest taki:

Listing.

Gen $N \sim \text{Poiss}(\Lambda(t))$;
for $i := 1$ **to** N **do** Gen $R_i \sim \text{U}(0, \Lambda(t))$; $\quad T_i := \Lambda^-(R_i)$

Pominęliśmy tu sortowanie skoków i założyliśmy, że symulacje ograniczamy do odcinka $[0, t]$.

Przykład 6.4 (Przerzedzanie)**.** To jest odpowiednik metody eliminacji. Załóżmy, że mamy dwie miary intensywności: μ o gęstości m i λ o gęstości l. To znaczy, że $\lambda(B) = \int_B l(x)\mathrm{d}x$ i $\mu(B) = \int_B m(x)\mathrm{d}x$ dla dowolnego zbioru $B \subseteq \mathcal{X}$. Załóżmy, że $l(x) \leqslant m(x)$ i przypuśćmy, że umiemy generować proces Poissona o intensywności μ. Niech $X_1, \ldots, X_N$ będą punktami tego procesu w zborze A o skończonej mierze μ (wiemy, że $N \sim \text{Poiss}(\mu(A))$. Punkt X_i *akceptujemy* z prawdopodobieństwem $l(X_i)/m(X_i)$ (pozostawiamy w zbiorze) lub odrzucamy (usuwamy ze zbioru) z prawdopodobieństwem $1-l(X_i)/m(X_i)$. Liczba pozostawionych punktów L ma rozkład $\text{Poiss}(\lambda(A))$, zaś każdy z tych punktów ma rozkład o gęstości $l(x)/\lambda(A)$, gdzie $\lambda(A) = \int_A l(x)\mathrm{d}x$. Te punkty tworzą proces Poissona z miarą intensywności λ.

Listing.

Gen $\mathbb{X} = \{X_1, \ldots, X_N\} \sim \text{Poiss}(\mu(\cdot))$;
for $i := 1$ **to** N Gen $U_i \sim \text{U}(0, 1)$;
  if $U_i > l(X_i)/m(X_i)$ **then** $\mathbb{X} := \mathbb{X} \setminus X_i$;
return $\mathbb{X} = \{X'_1, \ldots, X'_L\} \sim \text{Poiss}(\lambda(\cdot))$

Przykład 6.5 (Superpozycja)**.** To jest z kolei odpowiednik metody *kompozycji*. Metoda opiera się na następującym prostym fakcie. Jezeli $N_1(\cdot), \ldots, N_k(\cdot)$ są niezależnymi procesami Poissona z miarami intensywnosci odpowiednio $\mu_1(\cdot), \ldots, \mu_k(\cdot)$, to $N(\cdot) = \sum_i N_i(\cdot)$ jest procesem Poissona z intensywnością $\mu(\cdot) = \sum_i \mu_i(\cdot)$. Dodawanie należy tu rozumieć w dosłaowny sposób, to znaczy $N(A)$ jest określone jako $\sum_i N_i(A)$ dla każdego zbioru A. Jeśli utożsamimy procesy z losowymi zbiorami punktów to odpowiada temu operacja brania sumy mnogościowej (złączenia zbiorów). Niech $X_{j,1}, \ldots, X_{j,N_j}$ będą punktami j-tego procesu w zborze A o skończonej mierze μ_j.

Listing.

$\mathbb{X} = \emptyset$;
for $j := 1$ **to** k **do**
  begin
  Gen $\mathbb{X}_j = \{X_{j,1}, \ldots, X_{j,N_j}\} \sim \text{Poiss}(\mu_j(\cdot))$;
  $\mathbb{X} := \mathbb{X} \cup \mathbb{X}_j$;
  end
return $\mathbb{X} = \{X'_1, \ldots, X'_N\} \sim \text{Poiss}(\mu(\cdot))$
  { *mamy tu* $N = \sum_j N_j$ }

Wygodnie jest utożsamiać procesy Poissona z losowymi zbiorami punktów, jak uczyniliśmy w ostatnich przykładach (i mniej jawnie w wielu miejscach wcześniej). Te zbiory można rozumieć w zwykłym sensie, dodawać, odejmować tak jak w teorii mnogości pod warunkiem, że ich elementy się *nie powtarzają*. W praktyce mamy najczęściej do czynienia z intensywnościami, które mają gęstości „w zwykłym sensie", czyli względem miary Lebesgue'a. Wtedy, z prawdopodobieństwem 1, punkty procesu Poissona nie powtarzają się.

7. Symulowanie procesów stochastycznych II. Procesy Markowa

7.1. Czas dyskretny, przestrzeń dyskretna

Zacznijmy od najprostszej sytuacji. Rozważmy skończony zbiór $\mathcal{X}$, który będziemy nazywali *przestrzenią stanów*. Czasem wygodnie przyjąć, że $\mathcal{X} = \{1, 2, \ldots, d\}$. Jest to tylko umowne ponumerowanie stanów.

Definicja 7.1. (i) Ciąg $X_0, X_1, \ldots, X_n, \ldots$ zmiennych losowych o wartościach w $\mathcal{X}$ nazywamy **łańcuchem Markowa**, jeśli dla każdego $n = 1, 2, \ldots$ i dla każdego ciągu $x_0, x_1, \ldots, x_n, x_{n+1}$ punktów przestrzeni $\mathcal{X}$,

$$\begin{aligned}&\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0)\\ &= \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n),\end{aligned}$$

(o ile tylko $\mathbb{P}(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0) > 0$, czyli prawdopodobieństwo warunkowe w tym wzorze ma sens).

(ii) Łańcuch Markowa nazywamy **jednorodnym**, jeśli dla dowolnych stanów x i y i każdego n możemy napisać

$$\mathbb{P}(X_{n+1} = y | X_n = x) = P(x, y),$$

to znaczy prawdopodobieństwo warunkowe w powyższym wzorze zależy tylko od x i y, ale nie zależy od n.

Jeśli $X_n = x$, to mówimy, że łańcuch w chwili n znajduje się w stanie $x \in \mathcal{X}$. Warunek (i) w Definicji 7.1 znaczy tyle, że przyszła ewolucja łańcucha zależy od stanu obecnego, ale nie zależy od przeszłości. Lańcuch jest jednorodny, jeśli prawo ewolucji łańcucha nie zmienia się w czasie. W dalszym ciągu rozpatrywać będziemy głównie łańcuchy jednorodne. Macierz

$$P = (P(x,y))_{x,y\in\mathcal{X}} = \begin{pmatrix} P(1,1) & \cdots & P(1,y) & \cdots & P(1,d) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ P(x,1) & \cdots & P(x,y) & \cdots & P(x,d) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ P(d,1) & \cdots & P(d,y) & \cdots & P(d,d) \end{pmatrix}$$

nazywamy *macierzą* (prawdopodobieństw) *przejścia* łańcucha. Jest to macierz stochastyczna, to znaczy $P(x, y) \geqslant 0$ dla dowolnych stanów $x, y \in \mathcal{X}$ oraz $\sum_y P(x, y) = 1$ dla każdego $x \in \mathcal{X}$. Jeśli $q(x) = \mathbb{P}(X_0 = x)$, to wektor wierszowy

$$q^\top = (q(1), \ldots, q(x), \ldots, q(d))$$

nazywamy *rozkładem początkowym* łańcucha (oczywiście, $\sum_x q(x) = 1$). Jest jasne, że

$$\begin{aligned}&\mathbb{P}(X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n)\\ &= q(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n).\end{aligned} \tag{7.1}$$

Ten wzór określa jednoznacznie łączny rozkład prawdopodobieństwa zmiennych $X_0, X_1, \ldots, X_n, \ldots$ i może być przyjęty za definicję jednorodnego łańcucha Markowa. W tym sensie możemy utożsamić łańcuch z parą (q, P): łączny rozkład prawdopodobieństwa jest wyznaczony przez podanie rozkładu początkowego i macierzy przejścia.

Opiszemy teraz bardzo ogólną konstrukcję, która jest podstawą algorytmów generujących łańcuchy Markowa. Wyobraźmy sobie, jak zwykle, że mamy do dyspozycji ciąg $U = U_0, U_1, \ldots, U_n, \ldots$ „liczb losowych", produkowanych przez komputerowy generator, czyli z teoretycznego punktu widzenia ciąg *niezależnych zmiennych losowych o jednakowym rozkładzie*, $\mathrm{U}(0,1)$. Niech $\phi : [0,1] \to \mathcal{X}$ i $\psi : \mathcal{X} \times [0,1] \to \mathcal{X}$ będą takimi funkcjami, że

$$\begin{aligned} &\mathbb{P}\left(\psi(U) = x\right) = q(x) \quad \text{dla każdego } x \in \mathcal{X}, \\ &\mathbb{P}\left(\phi(x, U) = y\right) = P(x, y) \quad \text{dla dowolnych } x, y \in \mathcal{X}. \end{aligned} \tag{7.2}$$

Jeśli określimy $X_0, X_1, \ldots, X_n, \ldots$ rekurencyjnie w następujący sposób:

$$X_0 = \psi(U_0), \qquad X_{n+1} = \phi\left(X_n, U_{n+1}\right), \tag{7.3}$$

to jest jasne, że tak otrzymany ciąg zmiennych losowych jest łańcuchem Markowa z rozkładem początkowym q i macierzą przejścia P.

Na zakończenie tego podrozdziału zróbmy kilka prostych uwag.

— Wszystko, co w tym podrozdziale zostało powiedziane można niemal mechanicznie uogólnić na przypadek *nieskończonej ale przeliczalnej* przestrzeni $\mathcal{X}$.

— Można sobie wyobrażać, że dla ustalonego x, zbiór $\{u : \phi(x,u) = y\}$ jest *odcinkiem* długości $P(x,y)$, ale nie musimy tego żądać. Nie jest istotne, że zmienne U_i mają rozkład $\mathrm{U}(0,1)$. Moglibyśmy założyć, że są określone na dość dowolnej przestrzeni $\mathcal{U}$. Istotne jest, że te zmienne są niezależne, mają jednakowy rozkład i zachodzi wzór (7.2). W praktyce bardzo często do zrealizowania jednego „kroku" łańcucha Markowa generujemy więcej niż jedną „liczbę losową".

7.2. Czas dyskretny, przestrzeń ciągła

Przypadek ciągłej lub raczej *ogólnej* przestrzeni stanów jest w istocie równie prosty, tylko oznaczenia trzeba trochę zmienić i sumy zamienić na całki. Dla prostoty przyjmijmy, że $\mathcal{X} \subseteq \mathbb{R}^d$. Poniżej sformułujemy definicję w taki sposób, żeby podkreślić analogię do przypadku dyskretnego i pominiemy subtelności teoretyczne. Zwróćmy tylko uwagę, że prawdopodobieństwo warunkowe nie może tu być zdefiniowane tak elementarnie jak w Definicji 7.1, bo zdarzenie warunkujące może mieć prawdopodobieństwo zero.

Definicja 7.2. (i) Ciąg $X_0, X_1, \ldots, X_n, \ldots$ zmiennych losowych o wartościach w $\mathcal{X}$ nazywamy **łańcuchem Markowa**, jeśli dla każdego $n = 1, 2, \ldots$, dla każdego ciągu $x_0, x_1, \ldots, x_n$ punktów przestrzeni $\mathcal{X}$ oraz dowolnego (borelowskiego) zbioru $B \subseteq \mathcal{X}$,

$$\begin{aligned} &\mathbb{P}(X_{n+1} \in B | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0) \\ &= \mathbb{P}(X_{n+1} \in B | X_n = x_n), \end{aligned}$$

(dla prawie wszystkich punktów $(x_n, x_{n-1}, \ldots, x_1, x_0)$).

(ii) Łańcuch Markowa nazywamy **jednorodnym**, jeśli dla dowolnego zbioru $B \subseteq \mathcal{X}$ i każdego n możemy napisać

$$\mathbb{P}(X_{n+1} \in B | X_n = x) = P(x, B)$$

(dla prawie wszystkich punktów x).

Funkcja $P(\cdot,\cdot)$, której argumentami są *punkt x i zbiór B*, jest nazywana *jądrem przejścia*. Ważne dla nas jest to, że dla ustalonego $x \in \mathcal{X}$, jądro $P(x,\cdot)$ rozważane jako funkcja *zbioru* jest rozkładem prawdopodobieństwa. W wielu przykładach jest to rozkład zadany przez gęstość,

$$P(x,B) = \int_B p(x,y)\mathrm{d}y.$$

Wtedy odpowiednikiem wzoru (7.1) jest następujący wzór na łączną gęstość:

$$p(x_0,x_1,\ldots,x_{n-1},x_n) = q(x_0)p(x_0,x_1)\cdots p(x_{n-1},x_n), \tag{7.4}$$

gdzie $q(\cdot)$ jest gęstością rozkładu początkowego, zaś $p(x_i,x_{i+1}) = p(x_{i+1}|x_i)$ jest gęstością warunkową. Sformułowaliśmy Definicję 7.2 nieco ogólniej głównie dlatego, że w symulacjach ważną rolę odgrywają łańcuchy dla których prawdopodobieństwo przejścia nie ma gęstości. Przykładem może być łańcuch, który z niezerowym prawdopodobieństwem potrafi „stać w miejscu", to znaczy $P(x,\{x\}) =: \alpha(x) > 0$, i ma prawdopodobieństwo przejścia postaci, powiedzmy

$$P(x,B) = \alpha(x)\mathbb{I}(x \in B) + (1-\alpha(x))\int_B p(x,y)\mathrm{d}y.$$

Lańcuch Markowa na ogólnej przestrzeni stanów można rekurencyjnie generować w sposób opisany wzorem (7.3), to znaczy $X_0 = \psi(U_0)$ i $X_{n+1} = \phi(X_n,U_{n+1})$ dla ciągu i.i.d. $U_0,U_1,\ldots,U_n,\ldots$ z rozkładu $\mathrm{U}(0,1)$. Niech $Q(\cdot)$ oznacza rozkład początkowy. Funkcje ψ i ϕ muszą spełniać warunki

$$\begin{aligned} &\mathbb{P}(\psi(U) \in B) = Q(B) \quad \text{dla każdego (mierzalnego) } B \subseteq \mathcal{X},\\ &\mathbb{P}(\phi(x,U) \in B) = P(x,B) \quad \text{dla dowolnych } x \in \mathcal{X}, B \subseteq \mathcal{X}. \end{aligned} \tag{7.5}$$

Różnica między (7.2) i (7.5) jest tylko taka, że w ogólnej sytuacji rozkłady prawdopodobieństwa zmiennych losowych $\psi(U)$ i $\phi(x,U)$, czyli odpowiednio $Q(\cdot)$ i $P(x,\cdot)$ nie muszą być dyskretne. Ale można zastosować w zasadzie każdą metodę generacji zmiennych o zadanym rozkładzie, przy czym ten zadany rozkład zależy od x.

7.3. Czas ciągły, przestrzeń dyskretna

Rozważmy proces stochastyczny $(X(t), t \geqslant 0)$, czyli rodzinę zmiennych losowych o wartościach w *skończonej przestrzeni stanów* $\mathcal{X}$, indeksowaną przez parameter t, który interpretujemy jako *ciągły czas*. Pojedyncze stany oznaczamy przez $x,y,z,\ldots \in \mathcal{X}$ lub podobnie. Załóżmy, że trajektorie są prawostronnie ciągłymi funkcjami mającymi lewostronne granice (prawie na pewno).

Definicja 7.3. (i) Mówimy, że $X(t)$ jest **procesem Markowa**, jeśli dla dowolnych $x,y \in \mathcal{X}$ oraz $s,t \geqslant 0$,

$$\mathbb{P}(X(s+t) = y | X(s) = x, X(u), 0 \leqslant u < s) = \mathbb{P}(X(t+s) = y | X(s) = x).$$

(ii) Proces Markowa nazywamy **jednorodnym**, jeśli dla dowolnych stanów x i y i każdego t i s możemy napisać

$$\mathbb{P}(X(s+t) = y | X(s) = x) = P^t(x,y),$$

to znaczy prawdopodobieństwo warunkowe w powyższym wzorze zależy tylko od x, y i t, ale nie zależy od s.

Sens założenia (i) jest taki sam jak w Definicji 7.1. Żądamy mianowicie, żeby zmienna $X(t+s)$ byłą *warunowo niezależna* od zmiennych $X(u), 0 \leqslant u < s$ pod warunkiem zdarzenia $X(s) = x$. Oczywiście, $P^t(x, y)$ jest *prawdopodobieństwem przejścia w ciągu czasu* t. Rozkład początkowy oznaczymy przez $q(x) = \mathbb{P}(X(0) = x)$. Ponieważ $\mathcal{X}$ jest skończona, q może być przedstawiony jako wektor a P^t jako macierz.

Odpowiednikiem macierzy przejścia jest *macierz intensywności przejść* zdefiniowana następująco.

$$Q(x, y) = \lim_{t \to 0} \frac{1}{h} \left[P^h(x, y) - I(x, y) \right],$$

gdzie $I = P^0$ jest macierzą identycznościową. Można udowodnić, że granice istnieją. Tak więc $Q = \frac{\mathrm{d}}{\mathrm{d}t} P^t|_{t=0}$ i $P^h = I + hQ + o(h)$ przy $h \searrow 0$. Innymi słowy, $Q(x, y)$ jest rzeczywiście „intensywnością skoków z x do y" (na jednostkę czasu). Mamy wzór analogiczny do (6.3):

$$\begin{aligned} \mathbb{P}(X(t+h) = y | X(t) = x) &= hQ(x, y) + o(h) \text{ dla } x \neq y; \\ \mathbb{P}(X(t+h) = x | X(t) = x) &= 1 + hQ(x, x) + o(h), \quad h \searrow 0. \end{aligned}$$

Zauważmy, że mamy $\sum_y Q(x, y) = 0$. Wspomnijmy jeszcze mimochodem, że macierze przejścia wyrażają się przez macierz intensywności Q przy pomocy „macierzowej funkcji wykładniczej": $P^t = \exp[tQ]$. Dla uproszczenia notacji, niech $Q(x) = -Q(x, x) = \sum_{y \neq x} Q(x, y)$ oznacza "intensywność wszystkich skoków ze stanu x".

Z tego co zostało powiedziane łatwo się domyślić, że generowanie procesu Markowa można zorganizować podobnie jak dla procesu Poissona, poprzez *czasy skoków*. Niech $0 < T_1 < T_2 < \cdots < T_n < \cdots$ będą kolejnymi momentami skoków,

$$\begin{aligned} T_1 &= \inf \{ t > 0 : X(t) \neq X(0) \}; \\ T_{n+1} &= \inf \{ t > T_n : X(t) \neq X(T_n) \}. \end{aligned}$$

Jeśli $X(T_n) = x$, to czas oczekiwania na następny skok, $W_{n+1} = T_{n+1} - T_n$ zależy tylko od x i ma rozkład *wykładniczy* z parametrem $Q(x)$. W szczególności, mamy $\mathbb{E}(W_{n+1} | X(T_n) = x) = 1/Q(x)$. Jeśli obserwujemy proces w momentach skoków, czyli skupimy uwagę na ciągu $\hat{X}_n = X(T_n)$ to otrzymujemy łańcuch Markowa z czasem dyskretnym, nazywany czasem „szkieletem". Jego prawdopodobieństwa przejścia są takie:

$$\hat{P}(x, y) = \mathbb{P}(\hat{X}_{n+1} = y | \hat{X}_n = x) = \begin{cases} Q(x, y)/Q(x) & \text{jeśli } x \neq y; \\ 0 & \text{jeśli } x = y. \end{cases} \tag{7.6}$$

Oczywiście, cała trajektoria procesu $X(t)$ jest w pełni wyznaczona przez momenty skoków $T_1, \ldots, T_n, \ldots$ i kolejne stany łańcucha szkieletowego chain $X(0) = \hat{X}_0, \hat{X}_1, \ldots, \hat{X}_n, \ldots$. Po prostu, $X(t) = \hat{X}_n$ dla $T_n \leqslant t < T_{n+1}$. Następujący algorytm formalizuje opisany powyżej sposób generacji.

Listing.

```
Gen X̂_0 ~ q(·);
T_0 := 0;
for i := 1 to ∞ do
  begin
  Gen W_i ~ Ex(Q(X̂_{i-1}));
  T_i := T_{i-1} + W_i;
  Gen X̂_i ~ P̂(X̂_{i-1}, · ); { P̂ dane wzorem powyżej }
  end
```

Rzecz jasna, w praktyce trzeba ustalić sobie skończony horyzont czasowy h i przerwać symulacje gdy $T_i > h$. Zauważmy, że ten algorytm bez żadnych modyfikacji nadaje się do symulowania procesu Markowa na *nieskończoniej* ale *przeliczalnej* przestrzeni stanów, na przykład $\mathcal{X} = \{0, 1, 2, \dots\}$.

Czasami warto zmodyfikować algorytm w następujący sposób. Generowanie procesu Markowa można rozbić na dwie fazy: najpierw wygenerować „potencjalne czasy skoków" a następnie symulować „szkielet", który będzie skakał wyłącznie w poprzednio otrzymanych momentach (ale nie koniecznie we wszystkich spośród nich). Opiszemy dokładniej tę konstrukcję, która bazuje na własnościach procesu Poissona. Wyobraźmy sobie najpierw, że dla każdego stanu $x \in \mathcal{X}$ symulujemy niezależnie jednorodny proces Poissona $\mathbb{R}^x$ o punktach skoku $R_1^x < \dots < R_k^x < \dots$ przy czym ten proces ma intensywność $Q(x)$. Jeśli teraz, w drugiej fazie $\hat{X}_{i-1} = X(T_{i-1}) = x$, to za następny moment skoku wybierzemy najbliższy punkt procesu $\mathbb{R}^x$ na prawo od T_{i-1}, czyli $T_i = \min\{R_k^x : R_k^x > T_{i-1}\}$. Z własności procesu Poissona (patrz Stwierdzenie 6.2 i jego dowód) wynika, że zmienna $T_i - T_{i-1}$ ma rozkład wykładniczy z parametrem $Q(x)$ i metoda jest poprawna. Naprawdę nie ma nawet potrzeby generowania wszystkich procesów Poissona $\mathbb{R}^x$. Warto posłużyć się metodą *przerzedzania* i najpierw wygenerować jeden proces o dostatecznie dużej intensywności a następnie „w miarę potrzeby" go przerzedzać. Niech $\mathbb{R}^*$ będzie procesem Poissona o intensywności $Q^* \geqslant \max_x Q(x)$ i oznaczmy jego punkty skoków przez $\{R_1 < \dots < R_k < \dots\}$. Jeśli w drugiej fazie algorytmu mamy $X(R_{i-1}) = x$ w pewnym momencie $R_{i-1} \in \mathbb{R}^*$, to zaczynamy przerzedzać proces $\mathbb{R}^*$ w taki sposób, aby otrzymać proces o intensywności $Q(x) \leqslant Q^*$. Dla każdego punktu R_j, $j \geqslant i$ powinniśmy rzucić monetą i z prawdopodobieństwem $(1 - Q(x))/Q^*$ ten punkt usunąć. Ale jeśli usuniemy punkt R_i to znaczy, że w tym punkcie *nie ma skoku*, czyli $X(R_i) = x$. Jeśli punkt R_i zostawimy, to wykonujemy skok, losując kolejny stan z prawdopodobieństwem $\hat{P}(x, \cdot)$. W rezultacie, stan $X(R_i)$ jest wylosowany z rozkładu prawdopodobieństwa

$$\tilde{P}(x, y) = \begin{cases} Q(x, y)/Q^* & \text{jeśli } x \neq y; \\ 1 - Q(x)/Q^* & \text{jeśli } x = y. \end{cases} \tag{7.7}$$

Niech $\tilde{X}_n = X(R_i)$ będzie „nadmiarowym szkieletem" procesu. Jest to łańcuch Markowa, który (w odróżnieniu od „cieńszego szkieletu" $\hat{X}_n$) *może* w jednym kroku pozostać w tym samym stanie i ma prawdopodobieństwa przejścia $\mathbb{P}(\tilde{X}_n = y | \tilde{X}_{n-1} = x) = \tilde{P}(x, y)$.

Odpowiada temu następujący dwufazowy algorytm.

— **Faza I**

Listing.

```
Gen ℝ ~ Poiss(Q*) { proces Poissona }
```

— **Faza II**

Listing.

```
Gen X̃_0 ~ q(·);
for i := 1 to ∞ do Gen X̃_i ~ P̃(X̂_{i-1}, · ); { P̃ dane wzorem powyżej }
```

Algorytm jest bardzo podobny do metody wynalezionej przez Gillespie'go do symulowania wielowymiarowych procesów urodzin i śmierci (głównie w zastosowaniach do chemii). Te procesy mają najczęściej przestrzeń stanów postaci $\mathcal{X} = \{0, 1, 2, \dots\}^d$; stanem układu jest układ liczb $x = (x(1), \dots, x(d))$, gdzie $x(i)$ jest liczbą osobników (na przykład cząstek) typu i. Przestrzeń jest nieskończona, ale większa część rozważań przenosi się bez trudu. Pojawia się tylko problem

w ostatnim algorytmie jeśli $\max_x Q(x) = \infty$, czego nie można wykluczyć. Zeby sobie z tym poradzić, można wziąć Q^* dostatecznie duże, żeby „na ogół" wystarczało, a w mało prawdopodobnym przypadku zawitania do stanu x z $Q(x) > Q^*$ przerzucać się na pierwszy, jednofazowy algorym.

8. Algorytmy Monte Carlo I. Obliczanie całek

Duża część zastosowań metod MC wiąże się z obliczaniem całek lub sum. Typowe zadanie polega na obliczeniu wartości oczekiwanej

$$\theta = \mathbb{E}_\pi f(X) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x),$$

gdzie X jest zmienną losową o rozkładzie prawdopodobieństwa π na przestrzeni $\mathcal{X}$, zaś $f : \mathcal{X} \to \mathbb{R}$. Zazwyczaj $\mathcal{X}$ jest podzbiorem wielowymiarowej przestrzeni euklidesowej lub jest zbiorem skończonym ale bardzo licznym (na przykład zbiorem pewnych obiektów kombinatorycznych). Jeśli $\mathcal{X} \subseteq \mathbb{R}^d$ i rozkład π jest opisany przez gęstość p to całka określająca wartość oczekiwaną jest zwykłą całką Lebesgue'a. W tym przypadku zatem możemy napisać

$$\theta = \int_{\mathcal{X}} f(x)p(x)\mathrm{d}x.$$

Równie ważny w zastosowaniach jest przypadek *dyskretnej* przestrzeni $\mathcal{X}$. Jeśli $p(x) = \mathbb{P}(X = x)$, to

$$\theta = \sum_{x\in\mathcal{X}} f(x)p(x).$$

W dalszym ciągu tego rozdziału utożsamiamy rozkład π z funkcją p. Zwróćmy uwagę, że przedstawiamy tu θ jako wartość oczekiwaną tylko dla wygody oznaczeń; w istocie *każda* całka, suma, (a także prawdopodobieństwo zdarzenia losowego) jest wartością oczekiwaną.

Na pierwszy rzut oka nie widać czym polega problem! Sumowanie wykonuje każdy kalkulator, całki się sprawnie oblicza numerycznie. Ale nie zawsze. Metody MC przydają się w sytuacjach gdy spotyka się z następującymi trudnościami (lub przynajmniej którąś z nich).

— Przestrzeń $\mathcal{X}$ jest ogromna. To znaczy wymiar d jest bardzo duży lub skończona przestrzeń zawiera astronomicznie dużą liczbę punktów.
— Rozkład prawdopodobieństwa π jest „skupiony" w małej części ogromnej przestrzeni $\mathcal{X}$.
— Nie ma podstaw do zakładania, że funkcja f jest w jakimkolwiek sensie „gładka" (co jest warunkiem stosowania standardowych numerycznych metod całkowania).
— Gęstość p rozkładu π jest znana tylko z dokładnością do stałej normującej. Innymi słowy, umiemy obliczać $p'(x) = Zp(x)$ ale nie znamy stałej $Z = \int p'(x)\mathrm{d}x$. Czsem zadanie polega właśnie na obliczeni tej stałej (Z jest nazwane „funkcją podziału" lub sumą statystyczną).

Prosta metoda MC (po angielsku nazywana bardziej brutalnie: *Crude Monte Carlo*, czyli CMC) nasuwa się sama. Należy wygenerować n niezależnych zmiennych losowych $X_1, \ldots, X_n$ o jednakowym rozkładzie π i za estymator wartości oczekiwanej wziąć średnią z próbki,

$$\hat{\theta}_n = \hat{\theta}_n^{\mathrm{CMC}} = \sum_{i=1}^{n} f(X_i).$$

Mocne Prawo Wielkich Liczb gwarantuje, że $\hat{\theta}_n \to \theta$ prawie na pewno, gdy $n \to \infty$. W terminologii statystycznej, $\hat{\theta}_n$ jest *mocno zgodnym* estymatorem obliczanej wielkości. Zadanie wydaje się rozwiązane. Są jednak dwa zasadnicze kłopoty.

— Estymator $\hat{\theta}_n$ skonstruowany metodą CMC może się zbliżać do θ przerażająco wolno.
— Co robić, jeśli nie umiemy losować z rozkładu π?

8.1. Losowanie istotne

Zadziwiająco skutecznym sposobem na *oba* przedstawione wyżej kłopoty jest **losowanie istotne** (*Importance Sampling*, w skrócie IS). Przypuśćmy, że umiemy losować z rozkładu o gęstości q. Zauważmy, że

$$\theta = \int_{\mathcal{X}} \frac{p(x)}{q(x)} f(x) q(x) \mathrm{d}x = \mathbb{E}_q \frac{p(X)}{q(X)} f(X) = \mathbb{E}_q w(X) f(X),$$

gdzie $w(x) = p(x)/q(x)$. Piszemy tu wzory dla całek ale oczywiście dla sum jest tak samo. Niech $X_1, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o jednakowym rozkładzie g,

$$\hat{\theta}_n = \hat{\theta}_n^{\mathrm{IS1}} = \sum_{i=1}^{n} W_i f(X_i), \tag{8.1}$$

gdzie

$$W_i = w(X_i) = \frac{p(X_i)}{q(X_i)}$$

traktujemy jako *wagi* wylosowanych punktów X_i. Z tego co wyżej powiedzieliśmy wynika, że $\mathbb{E}_q \hat{\theta}_n = \theta$ oraz $\hat{\theta}_n \to \theta$ prawie na pewno, gdy $n \to \infty$. Mamy więc estymator *nieobciążony* i *zgodny*. Milcząco założyliśmy, że $q(X_i) > 0$ w każdym wylosowanym punkcie, czyli, że $\{x : p(x) > 0\} \subseteq \{x : q(x) > 0\}$. Z tym zwykle nie ma wielkiego kłopotu. Ponadto musimy założyć, że umiemy obliczać funkcję w w każdym wylosowanym punkcie. Jeśli znamy tylko $p'(x) = Zp(x)$ a nie znamy stałej Z to jest kłopot. Możemy tylko obliczyć $W_i' = p'(X_i)/q(X_i) = ZW_i$. Używamy zatem nieco innej postaci estymatora IS, mianowicie

$$\hat{\theta}_n = \hat{\theta}_n^{\mathrm{IS2}} = \frac{\sum_{i=1}^{n} W_i f(X_i)}{\sum_{i=1}^{n} W_i}. \tag{8.2}$$

Ten estymator wygląda dziwnie, bo w mianowniku występuje *estymator jedynki* ale chodzi o to, że we wzorze (8.2) w przeciwieństwie do (8.1) możemy zastąpić W_i przez W_i', bo nieznany czynnik Z skraca się. Możemy jeszcze zapisać nasz estymator w zgrabnej formie

$$\hat{\theta}_n^{\mathrm{IS2}} = \sum_{i=1}^{n} \tilde{W}_i f(X_i),$$

gdzie $\tilde{W}_i = W_i / \sum_j W_j$ są *unormowanymi wagami*. Trzeba jednak pamiętać, że to „unormowanie" polega na podzieleniu wag przez zmienną losową. W rezultacie estymator IS2 jest *obciążony*. Jest jednak mocno zgodny, bo licznik we wzorze (8.2) po podzieleniu przez n zmierza do θ a mianownik do 1, prawie na pewno. Estymator IS2 jest znacznie częściej używany niż IS1. Oprócz uniezależnienia się od stałej normującej, jest jeszcze inny powód. Okazuje się, że (pomimo obciążenia) estymator IS2 może być bardziej *efektywny*. Sens tego pojęcia wyjaśnimy w następnym podrozdziale. Zarówno dobór przykładów jak i ogólny tok rozważań jest zapożyczony z monografii Liu [15].

8.2. Efektywność estymatorów MC

Naturalne jest żądanie, aby estymator był mocno zgodny, $\hat{\theta}_n \to \theta$ prawie na pewno przy $n \to \infty$. Znaczy to, że zbliżymy się dowolnie blisko aproksymowanej wielkości, gdy tylko dostatecznie długo przedłużamy symulacje. Chcielibyśmy jednak wiedzieć coś konkretniejszego, oszacować jak szybko *błąd aproksymacji* $\hat{\theta}_n - \theta$ maleje do zera i jakie n wystarczy do osiągnięcia

wystarczającej dokładności. Przedstawimy teraz najprostsze i najczęściej stosowane w praktyce podejście, oparte na tzw. asymptotycznej wariancji i konstrukcji asymptotycznych przedziałów ufności. W typowej sytuacji estymator $\hat{\theta}_n$ ma następującą własność, nazywaną *asymptotyczną normalnością*:

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \to \mathrm{N}(0, \sigma^2), \qquad (n \to \infty)$$

w sensie zbieżności według rozkładu. W konsekwencji, dla ustalonej liczby $a > 0$,

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \leqslant \frac{z\sigma}{\sqrt{n}}\right) \to \Phi(z) - \Phi(-z) = 2\Phi(z) - 1, \qquad (n \to \infty),$$

gdzie Φ jest dystrybuantą rozkładu $\mathrm{N}(0,1)$. Często nie znamy *asymptotycznej wariancji* σ^2 ale umiemy skonstruować jej zgodny estymator $\hat{\sigma}_n^2$. Dla ustalonej „małej" dodatniej liczby α łatwo dobrać kwantyl rozkładu normalnego $z = z_{1-\alpha/2}$ tak, żeby $2\Phi(z) - 1 = 1 - \alpha$. Typowo $\alpha = 0.05$ i $z = z_{0.975} = 1.9600 \approx 2$. Otrzymujemy

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| \leqslant \frac{z_{1-\alpha/2}\hat{\sigma}_n}{\sqrt{n}}\right) \to 1 - \alpha. \quad n \to \infty.$$

Ten wzór interpretuje się w praktyce tak: estymator $\hat{\theta}_n$ ma błąd nie przekraczający $z\hat{\sigma}_n/\sqrt{n}$ z prawdopodobieństwem około $1 - \alpha$. W żargonie statystycznym $1 - \alpha$ jest nazywane asymptotycznym poziomem ufności.

Chociaż opisame powyżej podejście ma swoje słabe strony, to prowadzi do prostego kryterium porównywania estymatorów. Przypuśćmy, że mamy dwa estymatory $\hat{\theta}_n^{\mathrm{I}}$ i $\hat{\theta}_n^{\mathrm{II}}$, oba asymptotycznie normalne, o asymptotycznej wariancji σ_{I}^2 i σ_{II}^6 odpowiednio. Przypuśćmy dalej, że dla obliczenia pierwszego z tych estymatorów generujemy n_{I}^2 punktów, zaś dla drugiego n_{I}^2. Błędy obu estymatorów na tym mym poziomie istotności są ograniczone przez, odpowiednio, $z\sigma_{\mathrm{I}}/\sqrt{n_{\mathrm{I}}}$ i $z\sigma_{\mathrm{II}}/\sqrt{n_{\mathrm{II}}}$. Przyrównując te wyrażenia do siebie dochodzimy do wniosku, że oba estymatory osiągają podobną dokładność, jeśli $n_{\mathrm{I}}/n_{\mathrm{II}} = \sigma_{\mathrm{II}}^2/\sigma_{\mathrm{I}}^2$. Liczbę

$$\mathrm{eff}(\hat{\theta}_n^{\mathrm{I}}, \hat{\theta}_n^{\mathrm{II}}) = \frac{\sigma_{\mathrm{II}}^2}{\sigma_{\mathrm{I}}^2}$$

nazywamy *względną efektywnością* (asymptotyczną). Czasami dobrze jest wybrać za „naturalny punkt odniesienia" estymator CMC i zdefiniować *efektywność* estymatora $\hat{\theta}_n$ o asymptotycznej wariancji σ^2 jako

$$\mathrm{eff}(\hat{\theta}_n) = \mathrm{eff}(\hat{\theta}_n, \hat{\theta}_n^{\mathrm{CMC}}) = \frac{\sigma_{\mathrm{CMC}}^2}{\sigma^2}.$$

Mówi się też, że jeśli wygenerujemy próbkę n punktów i obliczymy $\hat{\theta}_n$ to „efektywna liczność próbki" (ESS, czyli *effective sample size*) jest $n/\mathrm{eff}(\hat{\theta}_n)$. Tyle bowiem należałoby wygenerować punktów stosując CMC, żeby osiągnąć podobną dokładność.

Asymptotyczna normalność prostego estymatora CMC wynika wprost z Centralnego Twierdzenia Granicznego (CTG). Istotnie, jeśli generujemy niezależnie $X_1, \ldots, X_n$ o jednakowym rozkładzie π, to

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n^{\mathrm{CMC}} - \theta\right) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f(X_i) - \theta) \\ &\to \mathrm{N}(0, \sigma_{\mathrm{CMC}}^2), \qquad (n \to \infty),\end{aligned}$$

gdzie

$$\sigma_{\mathrm{CMC}}^2 = \mathrm{Var}_p f(X) = \int (f(x) - \theta)^2 \, p(x)\mathrm{d}x.$$

Zupełnie podobnie, dla losowania istotnego w formie (8.1) otrzymujemy asymptotyczną normalność i przy tym

$$\sigma^2_{\mathrm{IS1}} = \mathrm{Var}_q w(X) f(X) = \mathbb{E}_q \frac{(f(X)p(X) - \theta q(X))^2}{q(X)^2} = \int \frac{(f(x)p(x) - \theta q(x))^2}{q(x)} \mathrm{d}x.$$

Z tego wzorku widać, że estymator może mieć wariancję *zero* jeśli $q(x) \propto f(x)p(x)$. Niestety, żeby obliczyć $q(x)$ potrzebna jest znajomość współczynnika proporcjonalności, który jest równy... θ. Nie wszystko jednak stracone. Pozostaje ważna reguła heurystyczna:

Gęstość $q(x)$ należy tak dobierać, aby jej „kształt" był zbliżony do funkcji $f(x)p(x)$.

Dla drugiej wersji losowania istotnego, (8.2), asymptotyczna normalność wynika z następujących rozważań:

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{IS2}}_n - \theta\right) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n (W_i f(X_i) - \theta W_i)}{\frac{1}{n} \sum_{i=1}^n W_i} \to \mathrm{N}(0, \sigma^2_{\mathrm{IS2}}), \qquad (n \to \infty),$$

gdzie

$$\begin{aligned}\sigma^2_{\mathrm{IS2}} &= \mathrm{Var}_q w(X)(f(X) - \theta) \\ &= \mathrm{Var}_q w(X) f(X) - 2\theta \mathrm{Cov}_q(w(X), w(X)f(X)) + \theta^2 \mathrm{Var}_q w(X) \\ &= \sigma^2_{\mathrm{IS1}} + \theta\left[-2\mathrm{Cov}_q(w(X), w(X)f(X)) + \theta \mathrm{Var}_q w(X)\right].\end{aligned}$$

Wyrażenie w kwadratowym nawiasie może być ujemne, jeśli jest duża dodatnia korelacja zmiennych $w(X)$ i $w(X)f(X)$. W tej sytuacji estymator IS2 jest lepszy od IS1. Okazuje się więc rzecz na pozór paradoksalna: dzielenie przez estymator jedynki może poprawić estymator. Poza tym oba estymatory IS1 i IS2 mogą mieć mniejszą (asymptotyczną) wariancję, niż CMC. Jeśli efektywność jest większa niż 100%, to używa się czasem określenia „estymator *superefektywny*".

Jeśli interesuje nas obliczenie wartości oczekiwanej dla wielu różnych funkcji f, to warto wprowadzić uniwersalny, niezależny of f wskaźnik efektywności. Niezłym takim wskaźnikiem jest $\mathrm{Var}_q w(X)$. Istotnie, „odchylenie χ^2" pomiędzy gęstością instrumentalną p i docelową p, jest określone poniższym wzorem:

$$\begin{aligned}\chi^2 = \chi^2(q,p) &= \int \frac{(q(x) - p(x))^2}{q(x)} \mathrm{d}x \\ &= \mathbb{E}_q \left(\frac{q(X) - p(X)}{q(X)}\right)^2 = \mathbb{E}_q \left(1 - w(X)\right)^2 \\ &= \mathrm{Var}_q w(X).\end{aligned}$$

Niestety, to „odchylenie" nie jest odległością, bo nie spełnia warunku symetrii. Niemniej widać, że małe wartości $\mathrm{Var}_q w(X)$ świadczą o bliskości q i p. Zauważmy, jeszcze, że $\mathrm{Var}_q w(X)$ jest wariancją asymptotyczną estymatora IS1 dla funkcji $f(x) = 1$.

Zgodnym estymatorem wielkości χ^2 jest

$$\hat{\chi}^2 = \frac{1}{n-1} \sum_{i=1}^n \frac{(\bar{W}_n - W_i)^2}{\bar{W}^2}.$$

Dzielenie przez $n-1$ (a nie n) wynika z tradycji. Dzielenie przez $\bar{W}^2$ powoduje, że estymator nie wymaga znajomości czynnika normującego gęstość p. Jest to w istocie estymator typu IS2.

Estymator typu IS1 jest równy po prostu $\sum(1-W_i)^2/n$, ale można go stosować tylko gdy znamy czynnik normujący.

Wprowadźmy teraz bardziej formalnie pojęcie ważonej próbki. Niech (X, W) będzie parą zmiennych losowych o wartościach w $\mathcal{X} \times [0, \infty[$. Jeśli rozkład brzegowy X ma gęstość q zaś docelowa gęstość jest postaci $p(x) = p'(x)/Z$ to żądamy aby

$$\mathbb{E}(W|X = x) = \frac{p'(x)}{q(x)}$$

dla prawie wszystkich x. Mówimy wtedy, że próbka jest *dopasowana* do rozkładu p. Idea jest taka jak w losowaniu istotnym IS2: dopasowanie gwarantuje, że dla każdej funkcji f mamy $\mathbb{E}_q f(X)W = Z\mathbb{E}_p f(X)$. Nie jest konieczne, aby W było deterministyczną funkcją X. Pozostawiamy sobie możliwość generowania wag losowo.

Stała normująca Z jest na ogół nieznana, a więc waga W jest określona z dokładnością do proporcjonalności. W poprzednich podrozdziałach dla podkreślenia opatrywaliśmy takie nieunormowane wagi znaczkiem „prim" ale teraz to pomijamy. Gdy generujemy ciąg ważonych zmiennych losowych $(X_i, W_i), \ldots, (X_n, W_n)$, to oczywiście wymaga się aby $\mathbb{E}(W_i|X_i = x) = Zp(x)/q(x)$, gdzie stała Z musi być jednakowa dla wszystkich $i = 1, \ldots, n$. Jeśli założy się niezależność par (X_i, W_i), to zgodność i asymptotyczną normalność estymatora $\hat{\theta}_n^{\mathrm{IS2}}$ danego równaniem (8.2) wykazuje się tak samo jak poprzednio. Asymptotyczna wariancja jest równa $Z^{-2}\mathrm{Var}_q W(f(X) - \theta)$; czynnik Z^{-2} pojawia się dlatego, że operujemy nieunormowanymi wagami.

Jest jednak dużo ciekawych zastosowań, w których punkty X_i są *zależne*. Wtedy nawet zgodność estymatora $\hat{\theta}_n^{\mathrm{IS2}}$ nie jest automatyczna. Asymptotyczna normalność może zachodzić z graniczną wariancją zupełnie inną niż w przypadku niezależnym lub nie zachodzić wcale.

8.3. Ważona eliminacja

Interesujące jest powiązanie idei eliminacji z ważeniem próbek. Czysta metoda eliminacji pracuje w sytuacji, gdy gęstość „instrumentalna" majoryzuje funkcję proporcjonalną do gęstości docelowej, czyli $Zp(x) = p'(x) \leqslant q(x)$. Jeśli ten warunek *nie jest* spełniony, to można naprawić odpowiednio *ważąc* wylosowaną próbkę. Dokładniej, algorytm ważonej eliminacji (*Rejection Control*, w skrócie RC) jest następujący.

Listing.

```
repeat
  Gen Y ~ q;
  W := p'(Y)/q(Y);
  if W ⩽ 1 then
    begin
    W := 1;
    Gen U ~ U(0,1)
    end;
until (W > 1 or U < W)
X := Y
```

Dowód poprawności algorytmu. Zauważmy, że prawdopodobieństwo akceptacji Y w tym algorytmie jest równe $a(Y)$, gdzie

$$a(x) = \frac{p'(x) \wedge q(x)}{q(x)} = \begin{cases} \dfrac{p'(x)}{q(x)} & \text{jeśli } p'(x) \leqslant q(x); \\ 1 & \text{jeśli } p'(x) > q(x). \end{cases}$$

Rozkład X na wyjściu algorytmu ma zatem gęstość $\propto p'(x) \wedge q(x)$. Waga $W = w(Y)$ jest więc „dopasowana" do rozkładu p w sensie zdefiniowanym w poprzednim podrozdziale, bo

$$w(x) = \frac{p'(x)}{p'(x) \wedge q(x)} = \begin{cases} 1 & \text{jeśli } p'(x) \leqslant q(x); \\ \dfrac{p'(x)}{q(x)} & \text{jeśli } p'(x) > q(x). \end{cases}$$

□

Zauważmy jeszcze, jak należy *modyfikować* wagi w RC, jeśli na wejściu mamy próbkę ważoną, dopasowaną do p. Jeżeli punkt Y, pochodzący z rozkładu q, ma na wejściu wagę $W_Y = p'(Y)/q(Y)$, to na wyjściu zaakceptowany punkt $X = Y$ ma rozkład $\propto p'(x) \wedge q(x)$ i powinien mieć wagę $W_X = p'(Y)/(p'(Y) \wedge q(Y))$. Widać stąd, że

$$W_Y = \frac{W_X}{a(Y)},$$

gdzie $a(\cdot)$ jest napisanym wyżej prawdopodobieństwem akceptacji w RC.

Przykład 8.1 (Nie-samo-przecinające się błądzenia)**.** Po angielsku nazywają się *Self Avoiding Walks*, w skrócie SAW. Niech $\mathbb{Z}^d$ będzie d-wymiarową kratą całkowitoliczbową. Mówimy, że ciąg $s = (0 = s_0, s_1, \ldots, s_k)$ punktów kraty jest SAW-em jeśli

— każde dwa kolejne punkty s_{i-1} i s_i sąsiadują ze sobą, czyli różnią się o ± 1 na dokładnie jednej współrzędnej,

— żadne dwa punkty nie zajmują tego samego miejsca, czyli $s_i \neq s_j$ dla $i \neq j$.

Zbiór wszystkich SAW-ów o k ogniwach w $\mathbb{Z}^d$ oznaczymy SAW_k^d, a dla d i k ustalonych w skrócie SAW. Przykład $s \in \mathrm{SAW}_{15}^2$ widać na rysunku. Natychmiast nasuwa się bardzo proste pytanie:

— Jak policzyć SAW-y, czyli obliczyć $L = L_{k,d} = |\mathrm{SAW}_k^d|$?

Zainteresujmy się teraz „losowo wybranym SAW-em". Rozumiemy przez to zmienną losową S o rozkładzie jednostajnym w zbiorze SAW_k^d, czyli taką, że $\mathbb{P}(S = s) = 1/L_{k,d}$ dla każdego $s \in \mathrm{SAW}_k^d$. W skrócie, $S \sim \mathrm{U(SAW)}$. Niech $\delta(s_k)$ oznacza odległość euklidesową końca SAW-a, czyli punktu s_k, od początku, czyli punktu 0. Na przykład dla łańcuszka widocznego na rysunku mamy $\delta(s_{15}) = \sqrt{5}$. Można zadać sobie pytanie, jaka jest średni kwadrat takiej odległości, czyli

— Jak obliczyć $\overline{\delta^2} = \overline{\delta^2_{d,k}} = \mathbb{E}\delta(S_k)^2$, przy założeniu, że $S \sim \mathrm{U(SAW)}$?

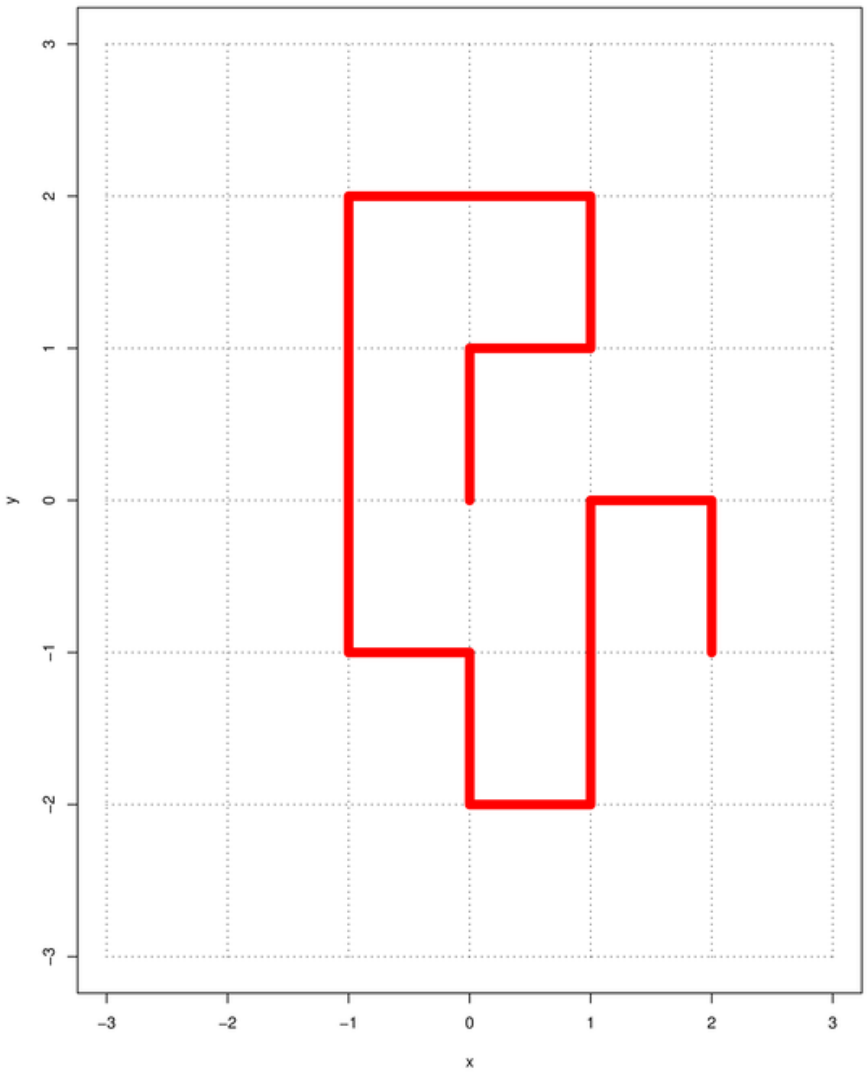
Można zastosować prostą metodę MC, czyli eliminację. Niech WALK_k^d oznacza zbiór wszystkich „błądzeń", czyli ciągów $s = (0 = s_0, s_1, \ldots, s_k)$ niekoniecznie spełniających warunek „$s_i \neq s_j$ dla $i \neq j$". Oczywiście $|\mathrm{WALK}_k^d| = (2d)^k$ i metoda generowania „losowego błądzenia" (z rozkładu U(WALK)) jest bardzo prosta: kolejno losujemy pojedyncze kroki, wybierając zawsze jedną z $2d$ możliwości. Żeby otrzymać „losowy SAW", stosujemy eliminację. Ten sposób pozwala w zasadzie estymować $\overline{\delta^2}$ (przez uśrednianie długości zaakceptowanych błądzeń) oraz SAW/WALK (przez zanotowanie frakcji akceptowanych błądzeń). Niestety, metoda jest bardzo nieefektywna, bo dla dużych k prawdopodobieństwo akceptacji szybko zbliża sie do zera.

Metoda „wzrostu" zaproponowana przez Rosenbluthów polega na losowaniu kolejnych kroków błądzenia spośród „dopuszczalnych punktów", to znaczy punktów wcześniej nie odwiedzonych. W każdym kroku, z wyjątkiem pierwszego mamy co najwyżej $2d - 1$ możliwości. W błądzeniu widocznym na rysunku kolejne kroki wybieraliśmy spośród:

$$4, 3, 3, 3, \quad 2, 3, 2, 2, \quad 3, 2, 3, 3, \quad 2, 1, 3, 2$$

możliwych. Nasz SAW został zatem wylosowany z prawdopodobieństwem

$$\frac{1}{4}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot \quad \frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot \quad \frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{3}\cdot\frac{1}{3}\cdot \quad \frac{1}{2}\cdot\frac{1}{1}\cdot\frac{1}{3}\cdot\frac{1}{2}$$



Rysunek 8.1. Przykład SAW-a.

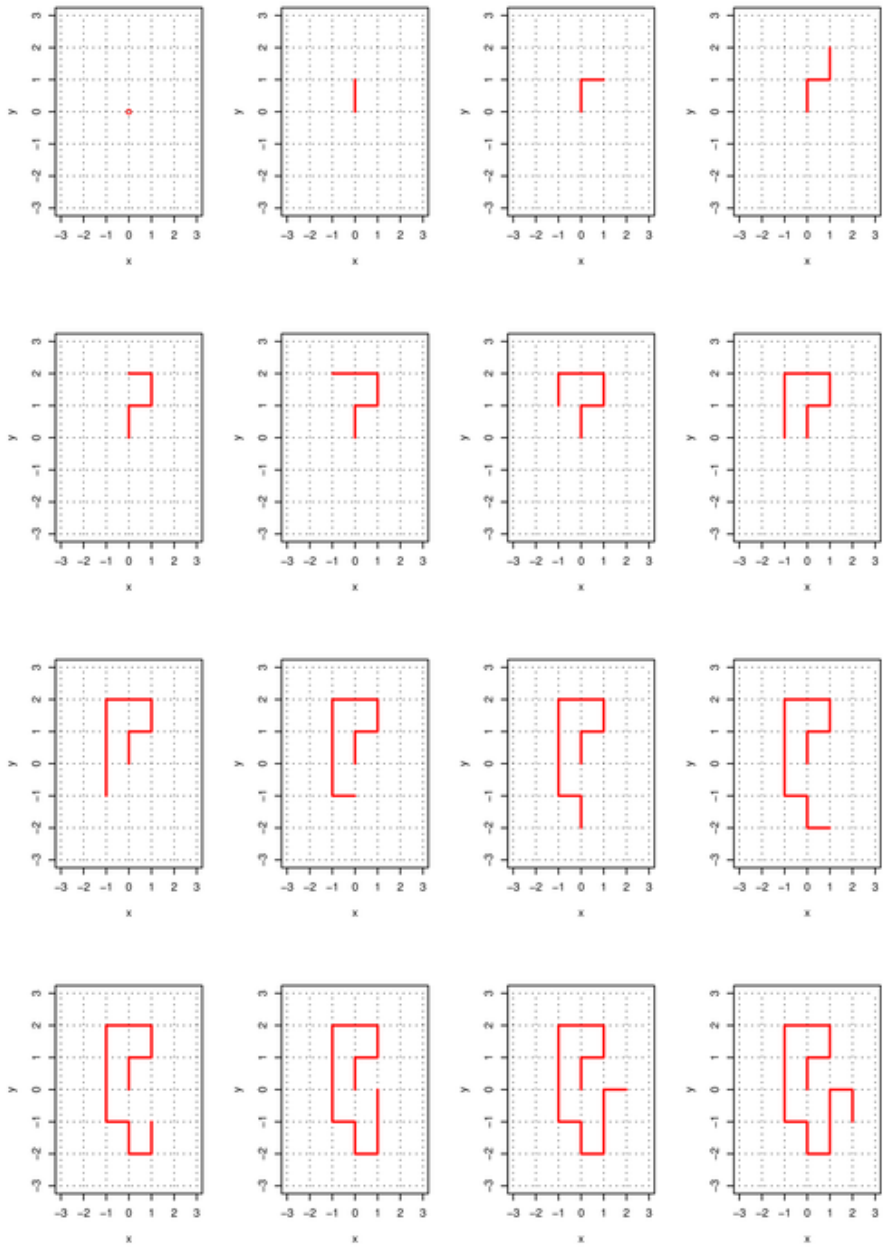
Powiedzmy ogólniej, że przy budowaniu SAW-a $s = (0 = s_0, s_1, \ldots, s_k)$ mamy kolejno

$$n_1 = 2d, n_2, \ldots, n_k$$

możliwości (nie jest przy tym wykluczone, że w pewnym kroku *nie mamy żadnej możliwości*, $n_i = 0$). Używając terminologii i oznaczeń związanych z losowaniem istotnym powiemy, że

$$q(s) = \frac{1}{n_1} \cdot \frac{1}{n_2}, \cdots \frac{1}{n_k}.$$

jest gęstością instrumentalną dla $s \in$ SAW, gęstość docelowa jest stała, równa $p(s) = 1/Z$, zatem funkcja podziału jest po prostu liczbą SAW-ów: $Z = |\text{SAW}|$. Wagi przypisujemy zgodnie ze wzorem $w(s) = n_1 \cdot n_2 \cdots n_k$ (jeśli wygenerowanie SAW-a się nie udało, $n_i = 0$ dla pewnego



Rysunek 8.2. Generowanie SAW-a metodą „wzrostu".

i, to waga jest zero). Niech teraz $S(1), \ldots, S(n)$ będą niezależnymi błądzeniami losowanymi metodą Rosenbluthów. Zgodnie z ogólnymi zasadami losowania istotnego,

$$\frac{\sum w(S(i))\delta(S(i))}{\sum w(S(i))} \quad \text{jest estymatorem } \overline{\delta^2},$$

$$\sum w(S(i)) \quad \text{jest estymatorem } |SAW|.$$

Należy przy tym *uwzględniać* w tych wzorach błądzenia o wadze zero, czyli „nieudane SAW-y".

Przykład 8.2 (Prawdopodobieństwo ruiny i wykładnicza zamiana miary)**.** Rozpatrzymy najprostszy model procesu opisującego straty i przychody w ubezpieczeniowej „teorii ryzyka". Niech

$Y_1, Y_2, \ldots$ będą zmiennymi losowymi oznaczającymi *straty netto* (straty − przychody) towarzystwa ubezpieczeniowego w kolejnych okresach czasu. Założymy (co jest dużym uproszczeniem), że te zmienne są niezależne i mają jednakowy rozkład o gęstości $p(y)$. Tak zwana „nadwyżka ubezpieczyciela" na koniec n-go roku jest równa

$$u - S_n = u - \sum_{i=1}^{n} Y_i,$$

gdzie u jest rezerwą początkową. Interesuje nas prawdopodobieństwo zdarzenia, polegającego na tym, że $u - S_n < 0$ dla pewnego n. Mówimy wtedy (znowu w dużym uproszczeniu) o „ruinie ubezpieczyciela". Wygodnie jest przyjąć następujące oznaczenia i konwencje. Zmienna losowa

$$R = \begin{cases} \min\{n : S_n > u\} & \text{jeśli takie } n \text{ istnieje;} \\ \infty & \text{jeśli } S_n \leqslant u \text{ dla wszystkich } n \end{cases}$$

oznacza czas oczekiwania na ruinę, przy czym jeśli ruina nigdy nie nastąpi to ten czas uznajemy za nieskończony. Przy takiej umowie, prawdopodobieństwo ruiny możemy zapisać jako $\psi = \mathbb{P}_p(R < \infty)$. Wskaźnik p przy symbolu wartości oczekiwanej przypomina, że chodzi tu o „oryginalny" proces, dla którego $Y_i \sim p$.

Obliczenie ψ analitycznie jest możliwe tylko w bardzo specjalnych przykładach. Motody numeryczne istnieją, ale też nie są łatwe. Pokażemy sposób obliczania ψ metodą Monte Carlo, który stanowi jeden z najpiękniejszych, klasycznych, przykładów losowania istotnego. Przyjmiemy bardzo rozsądne założenie, że $\mathbb{E}_p Y_i < 0$. *Funkcję tworzącą momenty*, która odpowiada gęstości p określamy wzorem

$$M_p(t) = \mathbb{E}_p \mathrm{e}^{tY_i} = \int_{-\infty}^{\infty} \mathrm{e}^{ty} p(y) \mathrm{d}y.$$

Założymy, że ta funkcja przyjmuje wartości skończone przynajmniej w pewnym otoczeniu zera i istnieje takie $r > 0$, że

$$M_p(t) = 1.$$

Liczba r jest nazywana *współczynnikiem dopasowania* i odgrywa tu kluczową rolę.

Metoda *wykładniczej zamiany miary* jest specjalnym przypadkiem losowania istotnego. Zeby określić instrumentalny rozkład prawdopodobieństwa, połóżmy

$$q(y) = \mathrm{e}^{ry} p(y).$$

Z definicji współczynnika dopasowania wynika, że q jest gęstością prawdopodobieństwa, to znaczy $\int q(y)\mathrm{d}y = 1$. Generuje się ciąg $Y_1, Y_2, \ldots$ jednakowo rozłożonych zmiennych losowych o gęstości q. Dla utworzonego w ten sposób „instrumentalnego procesu" używać będziemy symboli $\mathbb{P}_q$ i $\mathbb{E}_q$. Zauważmy, że

$$\begin{aligned} \mathbb{E}_q Y_i = \int y q(y) \mathrm{d}y &= \int y \mathrm{e}^{ry} p(y) \mathrm{d}y \\ &= \frac{\mathrm{d}}{\mathrm{d}r} \int \mathrm{e}^{ry} p(y) \mathrm{d}y \\ &= M_p'(r) > 0, \end{aligned}$$

ponieważ funkcja tworząca momenty $M_p(\cdot)$ jest wypukła, $M_p'(0) < 0$ i $M_p(0) = M_p(r) = 1$. Po zamianie miary, proces $u - S_n$ na mocy Prawa Wielkich Liczb zmierza prawie na pewno do minus nieskończoności, a zatem ruina następuje z prawdopodobieństwem jeden, $\mathbb{P}_q(R <$

$\infty) = 1$. Pokażemy, jak wyrazić prstwo ruiny dla oryginalnego procesu w terminach procesu instrumentalnego. Niech

$$\mathcal{R}_n = \{(y_1, \ldots, y_n) : y_1 \leqslant u, \ldots, y_1 + \cdots + y_{n-1} \leqslant u, y_1 + \cdots + y_{n-1} + y_n > u\},$$

Innymi słowy, zdarzenie $\{R = n\}$ zachodzi gdy $(Y_1, \ldots, Y_n) \in \mathcal{R}_n$. Mamy zatem

$$\begin{aligned}
\mathbb{P}_p(R = n) &= \int \cdots \int_{\mathcal{R}_n} p(y_1) \cdots p(y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n \\
&= \int \cdots \int_{\mathcal{R}_n} \mathrm{e}^{-ry_1} q(y_1) \cdots \mathrm{e}^{-ry_n} q(y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n \\
&= \int \cdots \int_{\mathcal{R}_n} \mathrm{e}^{-r(y_1 + \cdots + y_n)} q(y_1) \cdots q(y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n \\
&= \mathbb{E}_q \mathrm{e}^{rS_n} \mathbb{I}(R = n).
\end{aligned}$$

Weźmy sumę powyższych równości dla $n = 1, 2, \ldots$ i skorzystajmy z faktu, że $\sum_{n=1}^{\infty} \mathbb{P}_q(R = n) = 1$. Dochodzimy do wzoru

$$\mathbb{P}_p(R < \infty) = \mathbb{E}_q \exp\{-rS_R\} = \mathrm{e}^{-ru} \mathbb{E}_q \exp\{-r(S_R - u)\}. \tag{8.3}$$

Ten fakt jest podstawą algorytmu Monte Carlo:

Listing.

$\hat{\psi} := 0$; [$\hat{\psi}$ będzie estymatorem prawdopodobieństwa ruiny]

for $m := 1$ **to** k **do**
  begin
  $n := 0$; $S := 0$;
  repeat
    Gen $Y \sim q$; $S := S + Y$;
    until $S > u$;
  $Z := \exp\{-r(S - u)\}$; $\hat{\psi} := \hat{\psi} + Z$
  end
$\hat{\psi} := \mathrm{e}^{-ru}\hat{\psi}/k$

Algorytm jest prosty i efektywny. Trochę to zadziwiające, że w celu obliczenia prawdopodobieństwa ruiny generuje się proces dla którego ruina jest pewna. Po chwili zastanowienia można jednak zauważyć, że wykładnicza zamiana miary realizuje podstawową ideę losowania istotnego: rozkład instrumentalny „naśladuje" proces docelowy *ograniczony do zdarzenia* $\{R < \infty\}$.

Ciekawe, że wykładnicza zamiana miary nie tylko jest techniką Monte Carlo, ale jest też techniką *dowodzenia twierdzeń*! Aby się o tym przekonać, zauważmy, że „po drodze" udowodniliśmy nierówność $\psi \leqslant \mathrm{e}^{-ru}$ (wynika to z podstawowego wzoru (8.3) gdyż $\exp\{-r(S_R - u)\} \leqslant 1$). Jest to sławna nierówność Lundberga i wcale nie jest ona oczywista.

9. Algorytmy Monte Carlo II. Redukcja wariancji

W tym rozdziale omówię niektóre metody redukcji wariancji dla klasycznych algorytmów Monte Carlo, w których losujemy próbki niezależnie, z jednakowego rozkładu. Wiemy, że dla takich algorytmów wariancja estymatora zachowuje się jak const$/n$. Jedyne, co możemy zrobić – to konstruować takie algorytmy, dla których stała „const" jest możliwie mała. Najważniejszą z tych metod faktycznie już poznaliśmy: jest to *losowanie istotne*, omówione w Rozdziale 8. Odpowiedni wybór „rozkładu instrumentalnego" może zmniejszyć wariancję... setki tysięcy razy! Istnieje jeszcze kilka innych, bardzo skutecznych technik. Do podstawowych należą: losowanie warstwowe, metoda zmiennych kontrolnych, metoda zmiennych antytetycznych. Możliwe są niezliczone modyfikacje i kombinacje tych metod. Materiał zawarty w tym rozdziale jest w dużym stopniu zaczerpnięty z monografii Ripleya [18].

9.1. Losowanie warstwowe

Tak jak w poprzednim rozdziale, zedanie polega na obliczeniu wielkości

$$\theta = \mathbb{E}_\pi f(X) = \int_{\mathcal{X}} f(x)\pi(x)\mathrm{d}x,$$

gdzie rozkład prawdopodobieństwa i jego gęstość dla uproszczenia oznaczamy tym samym symbolem π. **Losowanie warstwowe** polega na tym, że rozbijamy przestrzeń X na sumę k rozłącznych podzbiorów (warstw),

$$\mathcal{X} = \bigcup_{h=1}^{k} A_h, \qquad A_h \cap A_g = \emptyset \quad (h \neq g),$$

i losujemy k niezależnych próbek, po jednej z każdej warstwy. Niech π_h będzie gęstostością rozkładu *warunkowego* zmiennej X przy $X \in A_h$, czyli

$$\pi_h(x) = \frac{\pi(x)}{p_h}\mathbb{I}(x \in A_h), \qquad \text{gdzie} \quad p_h = \pi(A_h) = \int_{A_h} \pi(x)\mathrm{d}x.$$

Widać natychmiast, że $\int_B \pi_h(x)\mathrm{d}x = \pi(X \in B | X \in A_h)$. Rozbijamy teraz całkę, którą chcemy obliczyć:

$$\theta = \sum_h p_h \int f(x)\pi_h(x)\mathrm{d}x.$$

Możemy użyć następującego estymatora warstwowego:

$$\hat{\theta}_n^{\mathrm{stra}} = \sum_h \frac{p_h}{n_h} \sum_{i=1}^{n_h} f(X_{hi}), \tag{9.1}$$

gdzie

$$X_{h1}, \ldots, X_{hn_h} \sim_{\text{i.i.d}} \pi_h,$$

jest próbką rozmiaru n_h wylosowaną z h-tej warstwy ($h = 1, \ldots, k$). Jest to estymator nieobciążony,

$$\mathrm{Var}\hat{\theta}_n^{\mathrm{stra}} = \sum_h \frac{p_h^2}{n_h}\sigma_h^2, \tag{9.2}$$

gdzie $\sigma_h^2 = \mathrm{Var}_\pi(f(X)|x \in A_h)$.

Porównajmy estymator warstwowy (9.1) ze „zwykłym" estymatorem

$$\hat{\theta}_n^{\mathrm{CMC}} = \frac{1}{n}\sum_{i=1}^n f(X_i).$$

opartym na jednej próbce

$$X_1, \ldots, X_n \sim_{\text{i.i.d}} \pi.$$

Oczywiście

$$\mathrm{Var}\hat{\theta}_n^{\mathrm{CMC}} = \frac{1}{n}\sigma^2,$$

gdzie $\sigma^2 = \mathrm{Var}_\pi f(X)$. Żeby porównanie było „uczciwe" rozważmy próbkę liczności $n = \sum_h n_h$. Jeśli decydujemy się na sumaryczną liczność próbki n, to możemy w różny sposób „rozdzielić" to n pomiędzy warstwami. To się nazywa *alokacja* próbki.

Stwierdzenie 9.1 (Alokacja proporcjonalna)**.** *Jeżeli $n_h = p_h n$ dla $n = 1, \ldots, k$, to*

$$\mathrm{Var}\hat{\theta}_n^{\mathrm{stra}} = \frac{1}{n}\sum_h p_h\sigma_h^2 \leqslant \frac{1}{n}\sigma^2 = \mathrm{Var}\hat{\theta}_n^{\mathrm{CMC}}.$$

Dowód. Wyrażenie na wariancję wynika znatychmiast z podstawienia $n_h = p_h n$ w ogólnym wzorze (9.2). Nierówność wynika z następującej tożsamości:

$$\sigma^2 = \sum_h p_h\sigma_h^2 + \sum_h p_h(\theta_h - \theta)^2, \tag{9.3}$$

gdzie $\theta_h = \int f(x)\pi_h(x)\mathrm{d}x = \mathbb{E}_\pi(f(X)|X \in A_h)$. Jest to klasyczny wzór na dekompozycję wariancji na „wariancję wewnątrz warstw" (pierwszy składnik w (9.3)) i „wariancję pomiędzy warstwami" (drugi składnik w (9.3)). Zdefiniujemy zmienną losową H jako „numer warstwy do której wpada $X \sim \pi$", czyli $\{H = h\} = \{X \in A_h\}$. Możemy teraz wzór (9.3) przepisać w dobrze znanej postaci

$$\mathrm{Var}\, f(X) = \mathbb{E}\mathrm{Var}(f(X)|H) + \mathrm{Var}\mathbb{E}(f(X)|H).$$

□

Wzór (9.3) podpowiada, jak dzielić przestrzeń na warstwy. Największy zysk w porównaniu z losowaniem nie-warstwowym jest wtedy, gdy „wariancja międzywarstwowa" jest dużo większa od „wewnątrzwarstwowej" Warstwy należy więc wybierać tak, żeby funkcja $\pi(c)f(x)$ była możliwie bliska stałej na każdym zbiorze A_h i różniła się jak najbardziej pomiędzy poszczególnymi zbiorami.

Stwierdzenie 9.1 pokazuje, że zawsze zyskujemy na losowaniu warstwowym, jeśli zastosujemy najprostszą, proporcjonalną alokację. Okazuje się, że nie jest to alokacja najlepsza. Jerzy Spława-Neyman odkrył prostą regułę wyznaczania alokacji optymalnej. Wychodzimy od wzoru (9.2) i staramy się zoptymalizować prawą stronę przy warunku $n = \sum_h n_h$. Poszukujemy zatem rozwiązania zadania

$$\sum_h \frac{p_h^2}{n_h}\sigma_h^2 = \min ! \qquad (\sum_h n_h - n = 0) \tag{9.4}$$

(względem zmiennych n_h). Zastosujmy metodę mnożników Lagrange'a. Szukamy minimum

$$\mathcal{L} = \sum_h \frac{p_h^2}{n_h}\sigma_h^2 + \lambda(\sum_h n_h - n) = \min ! \tag{9.5}$$

Obliczamy pochodną i przyrównujemy do zera:

$$\frac{\partial}{\partial n_h}\mathcal{L} = -\frac{p_h^2}{n_h^2}\sigma_h^2 + \lambda = 0. \tag{9.6}$$

Stąd natychmiast otrzymujemy rozwiązanie: $n_h \propto \sigma_h p_h$.

Stwierdzenie 9.2 (Alokacja optymalna, J. Neyman)**.** *Estymator warstwowy ma najmniejszą wariancję jeśli alokacja n losowanych punktów jest dana wzorem*

$$n_h = \frac{\sigma_h p_h}{\sum_g \sigma_g p_g}, \qquad (\ h = 1, \ldots, k\ ).$$

Zignorowaliśmy tutaj niewygodne wymaganie, że liczności próbek n_h muszą być całkowite. Gdyby to wziąć pod uwagę, rozwiązanie stałoby się skomplikowane, a zysk praktyczny z tego byłby znikomy. W praktyce rozwiązanie neymanowskie zaokrągla się do liczb całkowitych i koniec. Ważniejszy jest inny problem. Żeby wyznaczyć alokację optymalną, trzeba znać nie tylko prawdopodobieństwa p_h ale i wariancje warstwowe σ_h^2. W praktyce często opłaca się wylosować wstępne próbki, na podstawie których *estymuje* się wariancje σ_h^2. Dopiero potem alokuje się dużą, roboczą próbkę rozmiaru n, która jest wykorzystana do obliczania docelowej całki.

9.2. Zmienne kontrolne

Idea zmiennych kontrolnych polega na rozbiciu docelowej całki (wartości oczekiwanej) na dwa składniki, z których jeden umiemy obliczyć analitycznie. Metodę Monte Carlo stosujemy do drugiego składnika. Przedstawmy wielkość obliczaną w postaci

$$\theta = \mathbb{E}_\pi f(X) = \int_{\mathcal{X}} f(x)\pi(x)\mathrm{d}x = \int_{\mathcal{X}} [f(x) - k(x)]\pi(x)\mathrm{d}x + \int_{\mathcal{X}} k(x)\pi(x)\mathrm{d}x.$$

Dążymy do tego, żeby całka funkcji k była obliczona analitycznie (lub numerycznie) a różnica $f - k$ była możliwie bliska stałej, bo wtedy wariancja metody Monte Carlo jest mała. Funkcję k lub zmienną losową $k(X)$ nazywamy zmienną kontrolną.

Dla uproszczenia połóżmy $Y = f(X)$, gdzie $X \sim \pi$. Przypuśćmy, że zmiennej kontrolnej będziemy szukać pośród kombinacji liniowych funkcji $k_1, \ldots, k_d$ o znanych całkach. Niech $Z_j = k_j(X)$ i $Z = (Z_1, \ldots, Z_d)^\top$. Przy tym stale pamiętajmy, że $X \sim \pi$ i będziemy pomijać indeks π przy wartościach oczekiwanych i wariancjach. Zatem

$$k(x) = \sum_{j=1}^{d} \beta_j k_j(x).$$

Innymi słowy poszukujemy wektora współczynników $\beta = (\beta_1, \ldots, \beta_d)^\top$ który minimalizuje wariancję $\mathrm{Var}(Y - \beta^\top Z)$. Wykorzystamy następujący standardowy wynik z teorii regresji liniowej.

Stwierdzenie 9.3. *Niech Y i Z będą zmiennymi losowymi o skończonych drugich momentach, wymiaru odpowiednio* 1 *i* d*. Zakładamy dodatkowo odwracalność macierzy wariancji-kowariancji* $\mathrm{VAR}(Z)$*. Kowariancję pomiędzy Y i Z traktujemy jako wektor wierszowy i oznaczamy przez*

$\mathrm{COV}(Y,Z)$ *Spośród zmiennych losowych postaci* $Y-\beta^\top Z$*, najmniejszą wariancję otrzymujemy dla*

$$\beta_*^\top = \mathrm{COV}(Y,Z)\mathrm{VAR}(Z)^{-1}.$$

Ta najmniejsza wartość wariancji jest równa

$$\mathrm{Var}(Y-\beta_*^\top Z) = \mathrm{Var}Y - \mathrm{COV}(Y,Z)\mathrm{VAR}(Z)^{-1}\mathrm{COV}(Z,Y).$$

Przepiszmy tem wynik w bardziej sugestywnej formie. Niech $\mathrm{Var}Y=\sigma^2$. Można pokazać, że β_* maksymalizuje korelację pomiędzy Y i $\beta^\top Z$. Napiszmy

$$\varrho_{Y.Z} = \max_\beta \mathrm{corr}(Y,\beta^\top Z) = \mathrm{corr}(Y,\beta_*^\top Z) = \sqrt{\frac{\mathrm{COV}(Y,Z)\mathrm{VAR}(Z)^{-1}\mathrm{COV}(Z,Y)}{\mathrm{Var}Y}}.$$

Niech teraz $\hat\theta_n^{\mathrm{contr}}$ będzie estymatorem zmiennych kontrolnych. To znaczy, że losujemy próbkę $X_1,\dots,X_n\sim\pi$,

$$\hat\theta_n^{\mathrm{contr}} = \frac{1}{n}\sum(Y_i-\beta_*^\top Z_i)+\beta_*^\top\mu,$$

gdzie $Z_i^\top=(Z_{i1},\dots,Z_{in})=(f_1(X_i),\dots,f_d(X_i))$, zaś $\mu_j=\mathbb{E}f_j(X)$ są obliczone analitycznie lub numerycznie $\mu=(\mu_1,\dots,\mu_d)^\top$. Wariancja estymatora jest wyrażona wzorem

$$\mathrm{Var}\hat\theta_n^{\mathrm{contr}} = \frac{1}{n}\sigma^2(1-\varrho_{Y.Z}^2),$$

co należy porównać z wariancją σ^2/n „zwykłego" estymatora.

Zróbmy podobną uwagę, jak w poprzednim podrozdziale. Optymalny wybór współczynników regresji wymaga znajomości wariancji i kowariancji, których obliczenie może być ... (i jest zazwyczaj!) trudniejsze niż wyjściowe zadanie obliczenia wartości oczekiwanej. Niemniej, można najpierw wylosować *wstępną* próbkę, wyestymować potrzebne wariancje i kowariancje (nawet niezbyt dokładnie) po to, żeby dla dużej, *roboczej* próbki skonstruować dobre zmienne kontrolne.

Wspomnijmy na koniec, że dobieranie zmiennej kontrolnej metodą regresji liniowej nie jest jedynym sposobem. W konkretnych przykładach można spotkać najróżniejsze, bardzo pomysłowe konstrukcje, realizujące podstawową ideę zmiennych kontrolnych.

9.3. Zmienne antytetyczne

Przypuśćmy, że estymujemy wielkość $\theta=\mathbb{E}_\pi f(X)$. Jeśli mamy dwie zmienne losowe X i X' o jednakowym rozkładzie π ale nie zakładamy ich niezależności, to

$$\mathrm{Var}\frac{f(X)+f(X')}{2} = \frac{1}{2}\mathrm{Var}(X)\left[1+\mathrm{corr}(f(X),f(X'))\right] = \frac{1}{2}\sigma^2(1+\varrho).$$

Jeśli $\varrho=\mathrm{corr}(f(X),f(X'))<0$, to wariancja w powyższym wzorze jest *mniejsza* niż $\sigma^2/2$, czyli mniejsza niż w przypadku niezależności X i X'. To sugeruje, żeby zamiast losować *niezależnie* n zmiennych $X_1,\dots,X_n$, wygenerować *pary zmiennych ujemnie skorelowanych*. Załóżmy, że $n=2k$ i mamy k niezależnych par $(X_1,X_1'),\dots,(X_k,X_k')$, a więc łącznie n zmiennych. Porównajmy wariancję „zwykłego" estymatora $\hat\theta_n^{\mathrm{CMC}}$ i estymatora $\hat\theta_n^{\mathrm{ant}}$ wykorzystującego ujemne skorelowanie par:

$$\hat\theta_n^{\mathrm{CMC}} = \frac{1}{n}\sum_{i=1}^n f(X_i), \qquad \mathrm{Var}\hat\theta_n^{\mathrm{CMC}} = \frac{\sigma^2}{n}$$

$$\hat\theta_n^{\mathrm{ant}} = \frac{1}{n}\sum_{i=1}^{n/2}[f(X_i)=f(X_i')], \qquad \mathrm{Var}\hat\theta_n^{\mathrm{ant}} = \frac{\sigma^2}{n}(1+\varrho).$$

Wariancja estymatora $\hat{\theta}_n^{\text{ant}}$ jest tym mniejsza, im ϱ bliższe -1 (im bardziej ujemnie skorelowane są pary). Standardowym sposobem generowania ujemnie skorelowanych par jest odwracanie dystrybuanty z użyciem „odwróconych loczb losowych".

Stwierdzenie 9.4. *Jeśli $h :]0,1[\to \mathbb{R}$ jest funkcją monotoniczną różną od stałej, $\int_0^1 h(u)^2 \mathrm{d}u < \infty$ i $U \sim \mathrm{U}(0,1)$, to*

$$\mathrm{Cov}(h(U), h(1-U)) < 0.$$

Dowód. Bez straty ogólności załóżmy, że h jest niemalejąca. Niech $\mu = \mathbb{E}h(U) = \int_0^1 h(u)\mathrm{d}u$ i $t = \sup\{u : h(1-u) > \mu\}$. Latwo zauważyć, że $0 < t < 1$. Zauważmy, że

$$\begin{aligned}
\mathrm{Cov}(h(U), h(1-U)) &= \mathbb{E}h(U)[h(1-U) - \mu] \\
&= \int_0^1 h(u)[h(1-u) - \mu]\mathrm{d}u \\
&= \int_0^t h(u)[h(1-u) - \mu]\mathrm{d}u + \int_t^1 h(u)[h(1-u) - \mu]\mathrm{d}u \\
&< \int_0^t h(t)[h(1-u) - \mu]\mathrm{d}u + \int_t^1 h(t)[h(1-u) - \mu]\mathrm{d}u \\
&= h(t)\int_0^1 [h(1-u) - \mu]\mathrm{d}u = 0,
\end{aligned}$$

ponieważ dla $0 < u < t$ mamy $h(1-u) - \mu > 0$ i dla $t < u < 1$ mamy $h(1-u) - \mu \geqslant 0$. □

Przypomnijmy, że dla dowolnej dystrybuanty G określamy uogólnioną funkcję odwrotną G^- następującym wzorem:

$$G^-(u) = \inf\{x : G(x) \geqslant u\}.$$

Natychmiast wnioskujemy ze Stwierdzenia 9.4, że $\mathrm{Cov}(G^-(U), G^-(1-U)) < 0$. W ten sposób możemy produkować ujemnie skorelowane pary zmiennych o zadanej dystrybuancie. Okazuje się, że są to najbardziej ujemnie skorelowane pary. Jeśli $X \sim G$ i $X' \sim G$, to

$$\mathrm{Cov}(X, X') \geqslant \mathrm{Cov}(G^-(U), G^-(1-U)).$$

Powyższy fakt ma oczywiste znaczenie z punktu widzenia metod Monte Carlo. Udowodnimy ogólniejsze twierdzenie.

Twierdzenie 9.1. *Jeżeli $X \sim F$, $Y \sim G$ oraz $\mathbb{E}X^2 < \infty$, $\mathbb{E}Y^2 < \infty$ to*

$$\mathrm{Cov}(F^-(U), G^-(1-U)) \leqslant \mathrm{Cov}(X, Y) \leqslant \mathrm{Cov}(F^-(U), G^-(U)).$$

Twierdzenie 9.2 wynika z trzech poniższych faktów. Każdy z nich jest sam w sobie interesujący. Zaczniemy od sławnego wyniku Frecheta.

Twierdzenie 9.2 (Ograniczenia Frecheta)**.** *Jeżeli $\mathbb{P}(X \leqslant x) = F(x)$, $\mathbb{P}(Y \leqslant x) = G(y)$ oraz $\mathbb{P}(X \leqslant x, Y \leqslant y) = H(x,y)$ oznaczają, odpowiednio, dystrybuanty brzegowe oraz łączną dystrybuantę dwóch zmiennych losowych to*

$$\max(0, F(x) + G(y) - 1) \leqslant H(x,y) \leqslant \min(F(x), G(y)).$$

Istnieją rozkłady łączne o brzegowych F i G, dla których jest osiągane ograniczenie dolne i ograniczenie górne.

Dowód. Ograniczenie górne wynika z oczywistych nierówności

$$\mathbb{P}(X \leqslant x, Y \leqslant y) \leqslant \mathbb{P}(X \leqslant x) = F(x),$$
$$\mathbb{P}(X \leqslant x, Y \leqslant y) \leqslant \mathbb{P}(Y \leqslant y) = G(y).$$

Ograniczenie dolne jest równie proste:

$$\begin{aligned}\mathbb{P}(X \leqslant x, Y \leqslant y) &= \mathbb{P}(X \leqslant x) - \mathbb{P}(X \leqslant x, Y > y) \\ &\geqslant \mathbb{P}(X \leqslant x) - \mathbb{P}(Y > y) = F(x) - [1 - G(y)].\end{aligned}$$

□

Pozostała do pokazania osiągalność. Następujący lemat jest jednym z piękniejszych przykładów „symulacyjnego punktu widzenia" w rachunku prawdopodobieństwa.

Lemat 9.1. *Jeśli* $U \sim \mathrm{U}(0,1)$ *to*

$$\mathbb{P}(F^-(U) \leqslant x, G^-(U) \leqslant y) = \min(F(x), G(y));$$
$$\mathbb{P}(F^-(U) \leqslant x, G^-(1-U) \leqslant y) = \max(0, F(x) + G(y) - 1).$$

Dowód. Pierwsza równość jest oczywista:

$$\begin{aligned}\mathbb{P}(F^-(U) \leqslant x, G^-(U) \leqslant y) &= \mathbb{P}(U \leqslant F(x), U \leqslant G(y)) \\ &= \min(F(x), G(y)).\end{aligned}$$

Druga równość też jest oczywista:

$$\begin{aligned}\mathbb{P}(F^-(U) \leqslant x, G^-(1-U) \leqslant y) &= \mathbb{P}(U \leqslant F(x), 1 - U \leqslant G(y)) \\ &= \mathbb{P}(1 - G(y) \leqslant U \leqslant F(x)) \\ &= \max(0, F(x) - [1 - G(y)]).\end{aligned}$$

□

Głębokie twierdzenie Frecheta składa się więc z kilku dość oczywistych spostrzeżeń. Zeby udowodnić Twierdzenie 9.1 potrzeba jeszcze jednego ciekawego lematu.

Lemat 9.2. *Niech* $F(x)$, $G(y)$ *i* $H(x,y)$ *oznaczają, odpowiednio, dystrybuanty brzegowe oraz łączną dystrybuantę zmiennych losowych* X *i* Y. *Jeśli* $\mathbb{E}X^2 < \infty$, $\mathbb{E}Y^2 < \infty$ *to*

$$\mathrm{Cov}(X,Y) = \iint [H(x,y) - F(x)F(y)]\mathrm{d}x\mathrm{d}y.$$

Dowód. Niech (X_1, Y_1) i (X_2, Y_2) będą niezależnymi parami o jednakowym rozkładzie takim jak para (X, Y). Wtedy

$$\begin{aligned}2\mathrm{Cov}(X,Y) &= \mathbb{E}(X_1 - X_2)(Y_1 - Y_2) \\ &= \mathbb{E}\iint [\mathbb{I}(X_1 \leqslant x) - \mathbb{I}(X_2 \leqslant x)][\mathbb{I}(Y_1 \leqslant y) - \mathbb{I}(Y_2 \leqslant y)]\mathrm{d}x\mathrm{d}y \\ &= \iint \mathbb{E}[\mathbb{I}(X_1 \leqslant x) - \mathbb{I}(X_2 \leqslant x)][\mathbb{I}(Y_1 \leqslant y) - \mathbb{I}(Y_2 \leqslant y)]\mathrm{d}x\mathrm{d}y \\ &= \iint [\mathbb{P}(X_1 \leqslant x, Y_1 \leqslant y) + \mathbb{P}(X_2 \leqslant x, Y_2 \leqslant y) \\ &\qquad - \mathbb{P}(X_1 \leqslant x, Y_2 \leqslant y) - \mathbb{P}(X_2 \leqslant x, Y_1 \leqslant y)]\mathrm{d}x\mathrm{d}y \\ &= \iint [2\mathbb{P}(X \leqslant x, Y \leqslant y) - 2\mathbb{P}(X \leqslant x)\mathbb{P}(Y \leqslant y)]\mathrm{d}x\mathrm{d}y.\end{aligned}$$

□

10. Markowowskie Monte Carlo I. Wprowadzenie

10.1. Co to jest MCMC ?

W pewnych sytuacjach okazuje się, że wygenerowanie zmiennej losowej X z interesującego nas rozkładu prawdopodobieństwa π jest praktycznie niemożliwe. Wyobraźmy sobie, że π jest bardzo skomplikowanym rozkładem na „ogromnej", wielowymiarowej przestrzeni $\mathcal{X}$. Zazwyczaj ten rozkład jest dany poprzez podanie funkcji proporcjonalnej do gęstości, $p' \propto p$, ale bez znajomości stałej normującej $1/\int p'$. Przy tym „ogromna" przestrzeń $\mathcal{X}$ może być zbiorem skończonym (ale bardzo licznym) zaś rozkład π może być... jednostajny, o gęstości $p \propto 1$. Jeśli jednak nośnik rozkładu $\mathrm{U}(\mathcal{X})$ jest bardzo „skomplikowanym" zbiorem, to metody typu eliminacji/akceptacji mogą zawieść. Właśnie z takim przypadkiem spotkaliśmy się już we wstępnym Rozdziale 1, w Przykładzie 1.5. Naszkicowany tam algorytm jest przedstawicielem metod, którymi obecnie zajmiemy się dokładniej, czyli algorytmów MCMC. Ten skrót oznacza *Markov Chain Monte Carlo*, czyli po polsku algorytmy MC wykorzystujące łańcuchy Markowa. Podstawowa idea jest taka: jeśli nie umiemy generować zmiennej losowej X o rozkładzie π to zadowolimy się generowaniem ciągu zmiennych losowych $X_0, X_1, \ldots, X_n, \ldots$, który w pewnym sensie zbliża się, zmierza do rozkładu π. Podobnie jak w cytowanym przykładzie „plecakowym" ciąg X_n ma charakter „losowego błądzenia" po przestrzeni $\mathcal{X}$.

To wszystko co dotychczas powiedzieliśmy jest bardzo niekonkretne i wymaga uściślenia. W moich wykładach ścisłe przedstawienie tych zagadnień będzie możliwe tylko w ograniczonym zakresie. W obecnym rozdziale skupię się na głównych ideach MCMC. Następnie, w Rozdziałach 11 i 12 przedstawię podstawowe algorytmy i kilka motywujących przykładów, pokażę wyniki symulacji, ale niemal nic nie udowodnię. Sprobuję to częściowo naprawić w Rozdziałach 14 i 15 , które w całości będą poświęcone łańcuchom Markowa i algorytmom MCMC na skończonej przestrzeni $\mathcal{X}$. W tej uproszczonej sytuacji podam dowody (a przynajmniej szkice dowodów) podstawowych twierdzeń. Zastosowania MCMC w przypadku „ciągłej" przestrzeni $\mathcal{X}$ (powiedzmy, $\mathcal{X} \subseteq \mathbb{R}^d$) są przynajmniej równie ważne i w gruncie rzeczy bardzo podobne. Zobaczymy to na paru przykładach. Algorytmy MCMC pracujące na przestrzeni ciągłej są w zasadzie takie same jak te w przypadku przestrzeni skończonej, ale ich analiza robi się trudniejsza, wymaga więcej abstrakcyjnej matematyki – i w rezultacie wykracza poza zakres mojego skryptu. Ograniczę się w tej materii do kilku skromnych uwag. Bardziej systematyczne, a przy tym dość przystępne przedstawienie ogólnej teorii można znaleźć w [20] lub [17].

10.2. Lańcuchy Markowa

Klasyczne metody MCMC, jak sama nazwa wskazuje, opierają się na generowaniu łańcucha Markowa. Co prawda, rozwijają się obecnie bardziej wyrafinowane metody MCMC (zwane adaptacyjnymi), które wykorzystują procesy niejednorodne a nawet nie-markowowskie. Na razie ograniczymy się do rozpatrzenia sytuacji, gdy generowany ciąg zmiennych losowych X_n jest *jednorodnym łańcuchem Markowa* na przestrzeni $\mathcal{X}$. Jeśli $\mathcal{X}$ jest zbiorem skończonym, możemy

posługiwać się Definicją 7.1, w ogólnym przypadku – Definicją 7.2. Przypomnijmy oznaczenie *prawdopodobieństw przejścia* łańcucha: dla $x \in \mathcal{X}$ oraz $B \subseteq \mathcal{X}$,

$$\mathbb{P}(X_{n+1} \in B | X_n = x) = P(x, B).$$

W przypadku przestrzeni skończonej wygodniej posługiwać się *macierzą przejścia* o elementach

$$\mathbb{P}(X_{n+1} = y | X_n = x) = P(x, y).$$

Metoda generowania łańcuchów Markowa jest dość oczywista i sprowadza się do wzorów (7.2) w przypadku skończonym i (7.5) w przypadku ogólnym.

10.2.1. Rozkład stacjonarny

Niech π oznacza docelowy rozkład prawdopodobieństwa. Chcemy tak generować łańcuch X_n, czyli tak wybrać prawdopodobieństwa przejścia P, aby uzyskać zbieżność do rozkładu π. Spróbujemy uściślić co to znaczy, w jakim sensie rozumiemy zbieżność.

Definicja 10.1. Mówimy, że π jest **rozkładem stacjonarnym** (lub rozkładem równowagi) łańcucha Markowa o prawdopodobieństwach przejścia P, jeśli dla każdego (mierzalnego) zbioru $B \subseteq \mathcal{X}$ mamy

$$\pi(B) = \int_{\mathcal{X}} \pi(\mathrm{d}x) P(x, B).$$

W przypadku skończonej przestrzeni $\mathcal{X}$ równoważne sformułowanie jest takie: dla każdego stanu y,

$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x) P(x, y).$$

Jeśli rozkład początkowy jest rozkładem stacjonarnym, $\mathbb{P}(X_0 \in \cdot) = \pi(\cdot)$, to mamy $\mathbb{P}(X_n \in \cdot) = \pi(\cdot)$ dla każdego n. Co więcej, w takiej sytuacji łączny rozkład zmiennych $X_n, X_{n+1}, \ldots$ jest taki sam, jak rozkład zmiennych $X_0, X_1, \ldots$. Mówimy, że łańcuch jest w położeniu równowagi lub, ze jest *procesem stacjonarnym*. To uzasadnia nazwę rozkładu stacjonarnego. Oczywiście, łańcuchy generowane przez algorytmy MCMC nie są w stanie równowagi, bo z założenia nie umiemy wygenerować $X_0 \sim \pi$. Generujemy X_0 z pewnego innego rozkładu ξ, nazywanego rozkładem początkowym. Gdy zajdzie potrzeba, żeby uwidocznić zależność od rozkładu początkowego, będziemy używali oznaczeń $\mathbb{P}_\xi(\cdots)$ i $\mathbb{E}_\xi(\cdots)$. Przeważnie start jest po prostu deterministyczny, czyli ξ jest rozkładem skupionym w pewnym punkcie $x \in \mathcal{X}$. Piszemy wtedy $\mathbb{P}_x(\cdots)$ i $\mathbb{E}_x(\cdots)$.

Jeśli π jest rozkładem stacjonarnym, to przy pewnych dodatkowych założeniach uzyskuje się tak zwane *Słabe Twierdzenie Ergodyczne* (STE). Jego tezą jest zbieżność rozkładów prawdopodobieństwa zmiennych losowych X_n do π w następującym sensie: dla dowolnego (mierzalnego) zbioru $B \subseteq \mathcal{X}$ i dowolnego rozkładu początkowego ξ mamy

$$\mathbb{P}_\xi(X_n \in B) \to \pi(B) \qquad (n \to \infty). \tag{10.1}$$

Dla skończonej przestrzeni $\mathcal{X}$ równoważne jest stwierdzenie, że dla każdego $y \in \mathcal{X}$,

$$\mathbb{P}_\xi(X_n = y) \to \pi(y) \qquad (n \to \infty).$$

Pewną wersję STE dla skończonej przestrzeni $\mathcal{X}$ udowodnimy w następnym rozdziale (Twierdzenie 14.5). Na razie poprzestańmy na następującym prostym spostrzeżeniu.

Uwaga 10.1. Jeżeli zachodzi teza STE dla pewnego rozkładu granicznego π_∞, czyli $P^n(x, B) \to \pi_\infty(B)$ dla dowolnych $x \in \mathcal{X}$, $B \subseteq \mathcal{X}$, to π_∞ jest rozkładem stacjonarnym. W istocie, wystarczy przejść do granicy w równości

$$\begin{array}{ccc} P^{n+1}(x,B) & = & \int P^n(x,\mathrm{d}z)P(z,B) \\ \downarrow & & \downarrow \\ \pi_\infty(B) & & \int \pi_\infty(\mathrm{d}z)P(z,B). \end{array}$$

Co więcej, π_∞ jest *jedynym* rozkładem stacjonarnym.

Uwaga 10.2. W teorii łańcuchów Markowa rozważa się różne pojęcia zbieżności rozkładów. Zauważmy, że w powyżej przytoczonej tezie STE oraz w Uwadze 10.1 mamy do czynienia z silniejszym rodzajem zbieżności, niż poznana na rachunku prawdopodobieństwa zbieżność słaba (według rozkładu), oznaczana $\to_{\mathrm{d}}$.

10.2.2. Twierdzenia graniczne dla łańcuchów Markowa

Naszkicujemy podstawowe twierdzenia graniczne dla łańcuchów Markowa. Dokładniejsze sformułowania i dowody pojawią się w następnym rozdziale i będą ograniczone do przypadku skończonej przestrzeni $\mathcal{X}$. Bardziej dociekliwych Czytelników muszę odesłać do przeglądowych prac [20, 13].

Sformułujemy najpierw odpowiednik *Mocnego Prawa Wielkich Liczb* (PWL) dla łańcuchów Markowa. Rozważmy funkcję $f : \mathcal{X} \to \mathbb{R}$. Wartość oczekiwana funkcji f względem rozkładu π jest określona jako całka

$$\mathbb{E}_\pi f = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x).$$

Jeżeli rozkład π ma gęstość p względem miary Lebesgue'a to jest to „zwyczajna całka",

$$\mathbb{E}_\pi f = \int_{\mathcal{X}} f(x)p(x)\mathrm{d}x.$$

W przypadku *dyskretnej* przestrzeni $\mathcal{X}$ jest to suma

$$\mathbb{E}_\pi f = \sum_{x\in\mathcal{X}} f(x)\pi(x).$$

Jeśli załóżmy π jest rozkładem stacjonarnym łańcucha Markowa X_n, to możemy oczekiwać, że zachodzi zbieżność średnich do granicznej wartości oczekiwanej,

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \longrightarrow \mathbb{E}_\pi f \qquad (n \to \infty) \tag{10.2}$$

z prawdopodobieństwem 1. W istocie, można udowodnić (10.2) przy pewnych dodatkowych założeniach. Mówimy wtedy, że zachodzi PWL lub Mocne Twierdzenie Ergodyczne. Ze względu na zastosowania MCMC, wymagamy aby (10.2) zachodziło dla *dowolnego rozkładu początkowego* ξ (a nie dla $\xi = \pi$, czyli dla łańcucha stacjonarnego). Jedną z wersji PWL dla łańcuchów Markowa przedstawimy, wraz z pięknym i prostym dowodem, w Rozdziale 14 rozdziale (Twierdzenie 14.3).

Centralne Twierdzenie Graniczne (CTG) dla łańcuchów Markowa ma tezę następującej postaci. Dla dowolnego rozkładu początkowego ξ zachodzi zbieżność według rozkładu:

$$\frac{1}{\sqrt{n}}\left(\sum_{i=0}^{n-1}[f(X_i) - \mathbb{E}_\pi f]\right) \longrightarrow \mathrm{N}\big(0, \sigma^2_{\mathrm{as}}(f)\big) \quad (n \to \infty). \tag{10.3}$$

Liczba $\sigma^2_{\text{as}}(f)$, zwana asymptotyczną wariancją, nie zależy od rozkładu początkowego ξ, zależy zaś od macierzy przejścia P i funkcji f. Ponadto można udowodnić następujący fakt: dla dowolnego rozkładu początkowego ξ,

$$\frac{1}{n}\text{Var}_\xi\left(\sum_{i=0}^{n-1} f(X_i)\right) \longrightarrow \sigma^2_{\text{as}}(f). \tag{10.4}$$

Przy pewnych dodatkowych założeniach, asymptotyczną wariancję można wyrazić w terminach „stacjonarnych kowariancji" jak następuje. Mamy

$$\sigma^2_{\text{as}}(f) = \text{Var}_\pi f(X_0) + 2\sum_{n=1}^{\infty}\text{Cov}_\pi(f(X_0), f(X_n)), \tag{10.5}$$

gdzie Var_π i Cov_π oznaczają, oczywiście, wariancję i kowariancję obliczoną przy założeniu, że łańcuch jest stacjonarny. Niech

$$\begin{aligned}\sigma^2(f) =& \text{Var}_\pi f(X_0);\\ \sigma^2(f)\rho_n(f) =& \text{Cov}_\pi[f(X_0), f(X_n)],\end{aligned}$$

Przyjmijmy jeszcze, że $\rho_n(f) = \rho_{-n}(f)$. To określenie jest naturalne bo łańcuch stacjonarny można „przedłużyć wstecz". Wzór (10.5) można przepisać tak:

$$\sigma^2_{\text{as}}(f) = \sigma^2(f)\left(1 + 2\sum_{n=1}^{\infty}\rho_n(f)\right) = \sigma^2(f)\sum_{n=-\infty}^{\infty}\rho_n(f).$$

Później pojawi się kilka innych wzorów na asymptotyczną wariancję.

W tym miejscu chcę podkreślić różnicę między stacjonarną wariancją $\sigma^2(f) = \text{Var}_\pi f = \text{Var}_\pi f(X_n)$ i asymptotyczną wariancją $\sigma^2_{\text{as}}(f)$. W większości zastosowań kowariancje we wzorze (10.5) są dodatnie (zmienne losowe $f(X_0)$ i $f(X_k)$ są dodatnio skorelowane). W rezultacie $\sigma^2_{\text{as}}(f)$ jest dużo większa od $\sigma^2(f)$. To jest cena, którą płacimy za używanie łańcucha Markowa zamiast ciągu zmiennych niezależnych, jak w Rozdziale 8.

Zauważmy, że algorytmy Monte Carlo przeważnie mają za zadanie obliczyć pewną wartość oczekiwaną, a więc wielkość postaci $\theta = \mathbb{E}_\pi f$. Jeśli potrafimy generować łańcuch Markowa zbieżny do π, to naturalnym estymatorem $\mathbb{E}_\pi f$ jest

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i).$$

Możemy tezy PWL oraz CTG zapisać w skrócie tak:

$$\begin{aligned}\hat{\theta}_n &\longrightarrow_{\text{p.n.}} \theta \qquad (n\to\infty),\\ \sqrt{n}\left(\hat{\theta}_n - \theta\right) &\longrightarrow_{\text{d}} \text{N}(0, \sigma^2_{\text{as}}(f)), \quad (n\to\infty).\end{aligned}$$

PWL gwarantuje *zgodność* estymatora, a więc w pewnym sensie *poprawność* metody. Jest to, rzecz jasna, zaledwie wstęp do dokładniejszej analizy algorytmu. Zauważmy, że CTG może służyć do budowania asymptotycznych przedziałów ufności dla estymowanej wielkości θ: jeśli przyjmiemy poziom ufności $1-\alpha$ i dobierzemy odpowiedni kwantyl z rozkładu normalnego, to

$$\lim_{n\to\infty}\mathbb{P}\left(|\hat{\theta}_n - \theta| \leqslant \frac{z\sigma_{\text{as}}(f)}{\sqrt{n}}\right) = \Phi(z) - \Phi(-z) = 1-\alpha.$$

Oczywiście, $\Phi(z)$ oznacza to dystrybuantę rozkładu $\text{N}(0,1)$ i $\Phi(z) = 1-\alpha/2$. Potrzebne jest jeszcze oszacowanie asymptotycznej wariancji $\sigma^2_{\text{as}}(f)$, co nie jest wcale łatwe. Co więcej, CTG nie daje żadnej informacji o tym, dla jakich n przybliżenie rozkładem normalnym jest rozsądne.

Graniczne zachowanie wariancji estymatora $\hat{\theta}_n$ wyjaśnia wzór (10.4). Uzupełnijmy to (pomijając chwilowo uzasadnienie) opisem granicznego zachowania *obciążenia*: przy $n \to \infty$,

$$\begin{aligned} \mathrm{Var}_\xi(\hat{\theta}_n) &= \frac{1}{n}\sigma^2_{\mathrm{as}}(f) + o\left(\frac{1}{n}\right), \\ \mathbb{E}_\xi \hat{\theta}_n - \theta &= O\left(\frac{1}{n}\right). \end{aligned}$$

Oczywiście, naturalną miarą jakości estymatora jest *błąd średniokwadratowy* (BŚK). Ponieważ BŚK jest sumą wariancji i *kwadratu* obciążenia to, przynajmniej w granicy dla $n \to \infty$, wariancja ma dominujący wpływ, zaś obciążenie staje się zaniedbywalne:

$$\mathbb{E}_\xi \left(\hat{\theta}_n - \theta\right)^2 = \frac{1}{n}\sigma^2_{\mathrm{as}}(f) + o\left(\frac{1}{n}\right). \tag{10.6}$$

Powtóżmy jednak zastrzeżenie dotyczące wszystkich wyników zawartych w tym podrozdziale, Wzór (10.6) tylko sugeruje pewne przybliżenie interesującej nas wielkości.

Sławne i ważne twierdzenia graniczne sformułowane w tym podrozdziale nie są, niestety, całkowicie zadowalającym narzędziem analizy algorytmów Monte Carlo. Algorytmy wykorzystujące łańcuchy Markowa są użyteczne wtedy, gdy osiągają wystarczającą dokładność dla liczby kroków n *znikomo małej* w porównaniu z rozmiarem przestrzeni stanów. W przeciwnym przypadku można po prostu deterministycznie „przejrzeć wszystkie stany" i dokładnie obliczyć interesującą nas wielkość.

Niemniej, twierdzenia graniczne są interesujące z jakościowego punktu widzenia. Ponadto, ważne dla nas oszacowania dotyczące zachowania łańcucha w *skończonym czasie* można porównać z wielkościami granicznymi, aby ocenić ich jakość.

11. Markowowskie Monte Carlo II. Podstawowe algorytmy

11.1. Odwracalność

Najważniejsze algorytmy MCMC są oparte na idei odwracalności łańcucha Markowa.

Definicja 11.1. Łańcuch o jądrze P jest odwracalny względem rozkładu prawdopodobieństwa π, jeśli dla dowolnych $A, B \subseteq \mathcal{X}$ mamy

$$\int_A \pi(\mathrm{d}x)P(x,B) = \int_B \pi(\mathrm{d}y)P(y,A).$$

W skrócie,

$$\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \pi(\mathrm{d}y)P(y,\mathrm{d}x).$$

Odwracalność implikuje, że rozkład π jest stacjonarny. jest to dlatego ważne, że sprawdzanie odwracalności jest stosunkowo łatwe.

Twierdzenie 11.1. *Jeśli* $\pi(\mathrm{d}x)P(x,\mathrm{d}y) = \pi(\mathrm{d}y)P(y,\mathrm{d}x)$ *to* $\pi P = P$.

Dowód. $\pi P(B) = \int_{\mathcal{X}} \pi(\mathrm{d}x)P(x,B) = \int_B \pi(\mathrm{d}y)P(y,\mathcal{X}) = \int_B \pi(\mathrm{d}y) = \pi(B)$. □

11.2. Algorytm Metropolisa-Hastingsa

To jest pierwszy historycznie i wciąż najważniejszy algorytm MCMC. Zakładamy, że umiemy generować łańcuch Markowa z pewnym jądrem q. Pomysł Metropolisa polega na tym, żeby zmodyfikować ten łańcuch wprowadzając specjalnie dobraną regułę akceptacji w taki sposób, żeby wymusić zbieżność do zadanego rozkładu π. W dalszym ciągu systematycznie utożsamiamy rozkłady prawdopodobieństwa z ich gęstościami, aby nie mnożyć oznaczeń. Mamy zatem:
— Rozkład docelowy: $\pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$.
— Rozkład „propozycji": $q(x,\mathrm{d}y) = q(x,y)\mathrm{d}y$.
— Reguła akceptacji:

$$a(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \wedge 1.$$

Algorytm Metropolisa-Hastingsa (MH) interpretujemy jako „błądzenie losowe" zgodnie z jądrem przejścia q, zmodyfikowane poprzez odrzucanie niektórych ruchów, przy czym reguła akceptacji/odrzucania zależy w specjalny sposób od π. Pojedynczy krok algorytmu jest następujący.

Listing.

```
function KrokMH(x)
        Gen y ~ q(x,·); { propozycja }
        Gen U ~ U(0,1)
    if U > a(x,y)} then
      begin
      y := x; { ruch odrzucony z pr−stwem 1 − a(x,y) }
      end
KrokMH := y
```

Graficznie to można przedstawić w takiej postaci.

$$X_n = x \operatorname{ar}[d] \operatorname{ar}[dl]_{a(x,y)} y \sim q(x,\cdot) \operatorname{ar}[dr]^{1-a(x,y)} X_{n+1} = yX_{n+1} = x$$

Łańcuch Markowa powstaje zgodnie z następującym schematem:

Listing.

```
Gen X_0 ~ π_0; { start }
for n := 1 to ∞
    begin
    X_n := KrokMH(X_{n−1}) { krok }
    end
```

Oczywista jest analogia z podstawową metodą eliminacji. Zasadnicza różnica polega na tym, że w algorytmie MH nie „odrzucamy" zmiennej losowej, tylko „odrzucamy propozycje ruchu" – i stoimy w miejscu.

Jądro przejścia M-H jest następujące:

$$P(x,B) = \int_B \mathrm{d}y\, q(x,y)a(x,y) + \mathbb{I}(x \in B) \int_{\mathcal{X}} \mathrm{d}y\, q(x,y)[1-a(x,y)].$$

Dla przestrzeni skończonej, jądro łańcucha MH redukuje się do macierzy prawdopodobieństwprzejścia. Wzór jest w tym przypadku bardzo prosty: dla $x \neq y$,

$$P(x,y) = q(x,y)a(x,y).$$

Twierdzenie 11.2. *Jądro przejścia MH jest odwracalne względem π.*

Dowód. Ograniczmy się do przestrzeni skończonej, żeby nie komplikować oznaczeń. W ogólnym przypadku dowód jest w zasadzie taki sam, tylko napisy stają się mniej czytelne. Niech (bez straty ogólności)

$$a(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \leqslant 1, \qquad a(y,x) = 1.$$

Wtedy

$$\begin{aligned} \pi(x)P(x,y) &= \pi(x)q(x,y)a(x,y) \\ &= \pi(x)q(x,y)\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \\ &= \pi(y)q(y,x) \\ &= \pi(y)P(y,x) \quad \text{bo} \quad a(y,x) = 1. \end{aligned}$$

□

Uwagi historyczne:

— Metropolis w 1953 zaproponował algorytm, w którym zakłada się symetrię rozkładu propozycji, $q(x,y) = q(y,x)$. Warto zauważyć, że wtedy łańcuch odpowiadający q (błądzenie bez eliminacji ruchów) ma rozkład stacjonarny *jednostajny*. Reguła akceptacji przybiera postać

$$a(x,y) = \frac{\pi(y)}{\pi(x)} \wedge 1.$$

— Hastings w 1970 uogólnił rozważania na przypadek niesymetrycznego q.

Uwaga ważna:

— Algorytm MH wymaga znajomości gęstość π tylko z dokładnością do proporcjonalności, **bez stałej normującej**.

11.3. Próbnik Gibbsa

Drugim podstawowym algorytmem MCMC jest próbnik Gibbsa (PG) (*Gibbs Sampler*, GS). Załóżmy, że przestrzeń na której żyje docelowy rozkład π ma strukturę produktową: $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$. Przyjmijmy następujące oznaczenia:
— Jeśli $\mathcal{X} \ni x = (x_i)_{i=1}^d$ to $x_{-i} = (x_j)_{j\neq i}$: wektor z pominiętą i-tą współrzędną.
— Rozkład docelowy (gęstość): $\pi(\mathrm{d}x) = \pi(x)\mathrm{d}x$.
— Pełne rozkłady warunkowe (*full conditionals*):

$$\pi(x_i|x_{-i}) = \frac{\pi(x)}{\pi(x_{-i})}$$

Mały krok PG jest zmianą i-tej współrzędnej (wylosowaniem nowej wartości z rozkładu warunkowego):

$$\begin{aligned} x &= (x_1, \ldots, x_i, \ldots, x_d) \\ &\downarrow \\ \text{Gen } y_i &\sim \pi(\cdot|x_{-i}) \\ &\downarrow \\ Y &= (x_1, \ldots, y_i, \ldots, x_d). \end{aligned}$$

Prawdopodobieństwo przejścia małego kroku PG (w przypadku przestrzeni skończonej) jest takie:

$$P_i(x,y) = \pi(y_i|x_{-i})\mathbb{I}(x_{-i} = y_{-i}).$$

Twierdzenie 11.3. *Mały krok PG jest π-odwracalny.*

Dowód. Niech $x_{-i} = y_{-i}$. Wtedy

$$\begin{aligned} \pi(x)P_i(x,y) &= \pi(x)\pi(y_i|x_{-i}) \\ &= \pi(x_{-i})\pi(x_i|x_{-i})\pi(y_i|x_{-i}) \\ &= \pi(y_{-i})\pi(x_i|x_{-i})\pi(y_i|y_{-i}) \\ &= \pi(y)P_i(y,x). \end{aligned}$$

(skorzystaliśmy z symetrii). □

Oczywiście, trzeba jeszcze zadbać o to, żeby łańcuch generowany przez PG był nieprzywiedlny. Trzeba zmieniać wszystkie współrzędne, nie tylko jedną. Istnieją dwie zasadnicze odmiany próbnika Gibbsa, różniące się spoobem wyboru współrzędnych do zmiany.

— Losowy wybór współrzędnych, „LosPG".

— Systematyczny wybór współrzędnych, „SystemPG".

Losowy PG. Wybieramy współrzędną i-tą z prawdopodobieństwem $c(i)$. .

Listing.

function $LosPG(x)$
  Gen $i \sim c(\cdot)$;
  Gen $y_i := \pi(\cdot|x_{-i})$; { *zmieniamy $i-$tą współrzędną* }
  $y_{-i} := x_{-i}$; { *wszystkie inne współrzędne pozostawiamy bez zmian* }
$LosPG := y$

Jądro przejścia w „dużym" kroku losowego PG jest takie:

$$P = \sum_{i=1}^{d} c(i) P_i,$$

LosPG jest **odwracalny**.

Systematyczny PG. Współrzędne są zmieniane w porządku cyklicznym.

Listing.

function $SystemPG(x)$
  begin
  Gen $y_1 \sim \pi(\cdot|x_2, \ldots, x_d)$;
  Gen $y_2 \sim \pi(\cdot|y_1, x_3, \ldots, x_d)$;
  ...
  Gen $y_d \sim \pi(\cdot|y_1, \ldots, y_{d-1})$;
  $SystemPG := y$
  end

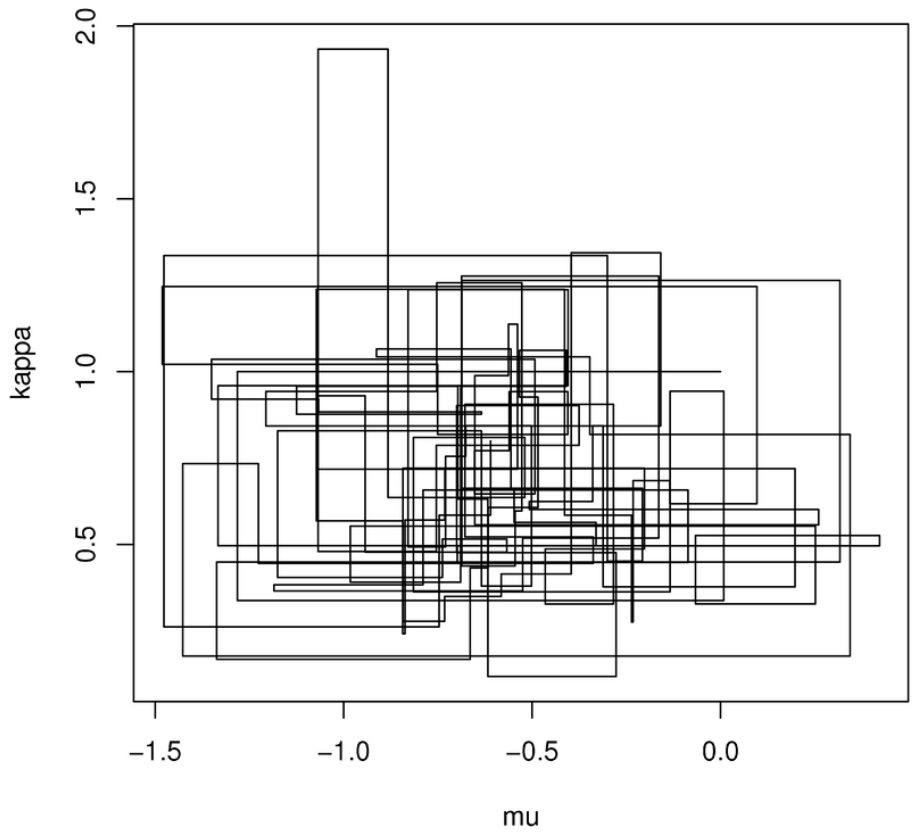
Jądro przejścia w „dużym" kroku systematycznego PGjest następujące.

$$P = P_1 P_2 \cdots P_d,$$

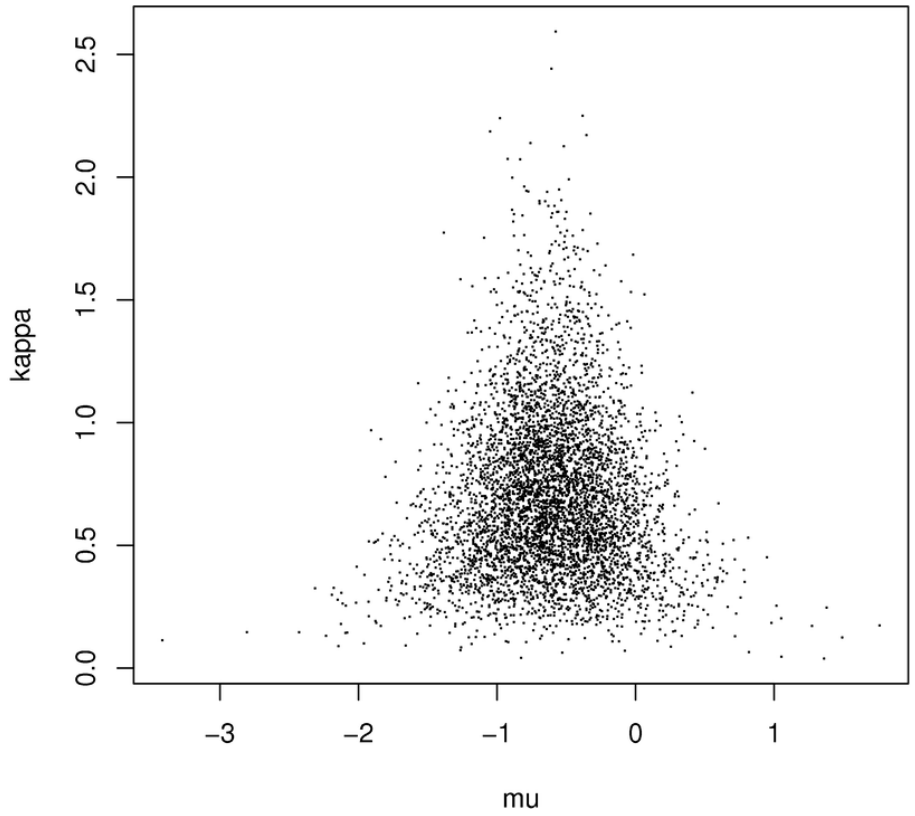
Systematyczny PG **nie jest odwracalny**. Ale jest π-stacjonarny, bo $\pi P_1 P_2 \cdots P_d = \pi$.

Przy projektowaniu konkretnych realizacji PG pojawia się szereg problemów, ważnych zarówno z praktycznego jak i teoretycznego punktu widzenia. Jak wybrać rozkład $c(\cdot)$ w losowym PG? Jest raczej jasne, że niektóre współrzędne powinny być zmieniane częściej, a inne rzadziej. Jak dobrać kolejność współrzędnych w systematycznym PG? Czy ta kolejność ma wpływ na tempo zbieżności łańcucha. Wreszcie, w wielu przykładach można zmieniać całe „bloki" współrzędnych na raz.

Systematyczny PG jest uważany za bardziej efektywny i częściej stosowany w praktyce. Z drugiej strony jest trudniejszy do analizy teoretycznej, niż losowy PG.



Rysunek 11.1. Trajektoria próbnika Gibbsa w przestrzeni dwuwymiarowej.



Rysunek 11.2. Chmurka punktów wygenerowanych przez próbnik Gibbsa.

12. Markowowskie Monte Carlo III. Przykłady zastosowań

12.1. Statystyka bayesowska

Algorytmy MCMC zrewolucjonizowały statystykę bayesowską. Stworzyły możliwość obliczania (w przybliżeniu) rozkładów *a posteriori* w sytuacji, gdy dokładne, analityczne wyrażenia są niedostępne. W ten sposób statystycy uwolnili się od konieczności używania nadmiernie uproszczonych modeli. Zaczęli śmiało budować modele coraz bardziej realistyczne, zwykle o strukturze hierarchicznej. Przedstawię to na jednym dość typowym przykładzie, opartym na pracy [21]. Inne przykłady i doskonały wstęp do tematyki zastosowań MCMC można znaleźć w pracy Geyera [7].

12.1.1. Hierarchiczny model klasyfikacji

Przykład 12.1 (Statystyka małych obszarów)**.** Zacznijmy od opisania problemu tak zwanych „małych obszarów", który jest dość ważny w dziedzinie badań reprezentacyjnych, czyli w tak zwanej „statystyce oficjalnej". Wyobraźmy sobie, że w celu zbadania kondycji przedsiębiorstw losuje się próbkę, która liczy (powiedzmy) 3500 przedsiębiorstw z całego kraju. Na podstawie tej losowej próbki można dość wiarygodnie szacować (estymować) pewne parametry opisujące populację przedsiębiorstw w kraju. Czy jednak można z rozsądną dokładnością oszacować sprzedaż w powiecie garwolińskim? W Polsce mamy ponad 350 powiatów. Na jeden powiat przypada średnio 10 przedsiębiorstw wybranych do próbki. *Małe obszary* to pod-populacje w których rozmiar próbki nie jest wystarczający, aby zastosować „zwykłe" estymatory (średnie z próbki). Podejście bayesowskie pozwala „pożyczać informację" z innych obszarów. Zakłada się, że z każdym małym obszarem związany jest nieznany parametr, który staramy się estymować. Obserwacje pochodzące z określonego obszaru mają rozkład prawdopodobieństwa zależny od odpowiadającego temu obszarowi parameru. Parametry, zgodnie z filozofią bayesowską, traktuje się jak zmienne losowe. W najprostszej wersji taki model jest zbudowany w następujący sposób.

Model bayesowski

— $y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$ – badana cecha dla j-tej wylosowanej jednostki i-tego obszaru, $(j = 1, \ldots, n_i)$, $(i = 1, \ldots, k)$,
— $\theta_i \sim \mathrm{N}(\mu, \upsilon^2)$ – interesująca nas średnia w i-tym obszarze,
— μ – średnia w całej populacji.

Ciekawe, że ten sam model pojawia się w różnych innych zastosowanianich, na przykład w matematyce ubezpieczeniowej. Przytoczymy klasyczny rezultat dotyczący tego modelu, aby wyjaśnić na czym polega wspomniane „pożyczanie informacji".

Estymator bayesowski

W modelu przedstawionym powyżej, łatwo obliczyć estymator bayesowski (przy kwadratowej funkcji straty), czyli wartość oczekiwaną *a posteriori*. Następujący wzór jest bardzo dobrze znany specjalistom od małych obszarów i aktuariuszom.

$$\hat{\theta}_i = \mathbb{E}(\theta_i|y) = z_i\bar{y}_i + (1-z_i)\mu, \qquad z_i = \frac{n_i\upsilon^2}{n_i\upsilon^2+\sigma^2}.$$

Estymator bayesowski dla i-tego obszaru jest średnią ważoną $\bar{y}$ (estymatora opartego na danych z tego obszaru) i wielkości μ, która opisuje całą populację, a nie tylko i-ty obszar. Niestety, proste estymator napisany powyżej zależy od parametrów μ, σ i υ, które w praktyce są nieznane i które trzeba estymować. Konsekwentnie bayesowskie podejście polega na traktowaniu również tych parametrów jako zmiennych losowych, czyli nałożeniu na nie rozkładów a priori. Powstaje w ten sposób model hierarchiczny.

Hierarchiczny model bayesowski

Uzupełnijmy rozpatrywany powyżej model, dobudowując „wyższe piętra" hierarchii. potraktujemy mianowicie parametry rozkładów *a priori*: μ, σ i υ jako zmienne losowe i wyspecyfikujemy ich rozkłady *a priori*.

— $y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$,
— $\theta_i \sim \mathrm{N}(\mu, \upsilon^2)$,
— $\mu \sim \mathrm{N}(m, \tau^2)$,
— $\sigma^{-2} \sim \mathrm{Gamma}(p, \lambda)$,
— $\upsilon^{-2} \sim \mathrm{Gamma}(q, \kappa)$.

Zakładamy przy tym, że μ, σ i υ są *a priori* niezależne (niestety, są one zależne *a posteriori*). Na szczycie hierarchii mamy „hiperparametry" m, τ, p, λ, q, κ, o których musimy założyć, ze są znanymi liczbami.

Łączny rozkład prawdopodobieństwa wszystkich zmiennych losowych w modelu ma postać

$$p(y, \theta, \mu, \sigma^{-2}, \upsilon^{-2}) = p(y|\theta, \sigma^{-2})p(\theta|\mu, \upsilon^{-2})p(\mu)p(\sigma^{-2})p(\upsilon^{-2}).$$

We wzorze powyżej i w dalej traktujemy (trochę nieformalnie) σ^{-2} i υ^{-2} jako pojedyncze symbole nowych zmiennych, żeby nie mnożyć oznaczeń. Rozkład prawdopodobieństwa *a posteriori* jest więc taki:

$$p(\theta, \mu, \sigma^{-2}, \upsilon^{-2}|y) = \frac{p(y, \theta, \mu, \sigma^{-2}, \upsilon^{-2})}{p(y)}.$$

To jest rozkład „docelowy" π, na przestrzeni $\mathcal{X} = \mathbb{R}^{k+3}$, ze nieznaną stałą normującą $1/p(y)$. Choć wygląda na papierze dość prosto, ale obliczenie rozkładów brzegowych, wartości oczekiwanych i innych charakterystyk jest, łagodnie mówiąc, trudne.

12.1.2. Próbnik Gibbsa w modelu hierarchicznym

Rozkłady warunkowe poszczególnych współrzędnych są proste i łatwe do generowania. Można te rozkłady „odczytać" uważnie patrząc na rozkład łączny:

$$\begin{aligned} p(\theta,\mu,\upsilon^{-2},\sigma^{-2}|y) \propto (\sigma^{-2})^{n/2}\exp\left\{-\frac{\sigma^{-2}}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2\right\} \\ \cdot(\upsilon^{-2})^{k/2}\exp\left\{-\frac{\upsilon^{-2}}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2\right\} \\ \cdot\exp\left\{-\frac{\tau^{-2}}{2}(\mu-m)^2\right\} \\ \cdot(\sigma^{-2})^{q-1}\exp\{-\kappa\sigma^{-2}\} \\ \cdot(\upsilon^{-2})^{p-1}\exp\{-\lambda\upsilon^{-2}\}. \end{aligned}$$

Dla ustalenia uwagi zajmijmy się rozkładem warunkowym zmiennej υ^{-2}. Kolorem niebieskim oznaczyliśmy te czynniki łącznej gęstości, które zawierają υ^{-2}. Pozostałe, czarne czynniki traktujemy jako stałe. Stąd widać, jak wygląda rozkład warunkowy υ^{-2}„ przynajmniej z dokładnością do proporcjonalności:

$$\begin{aligned} p(\upsilon^{-2}|y,\theta,\mu,\sigma^{-2}) \propto (\upsilon^{-2})^{k/2+p-1} \\ \cdot\exp\left\{-\left(\frac{1}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2+\lambda\right)\upsilon^{-2}\right\}. \end{aligned}$$

Jest to zatem rozkład Gamma$(k/2+p, \sum_{i=1}^{k}(\theta_i-\mu)^2/2+\lambda)$. Zupełnie podobnie rozpoznajemy inne (pełne) rozkłady warunkowe:

$$\begin{aligned} \upsilon^{-2}|y,\theta,\mu,\sigma^{-2} &\sim \text{Gamma}\left(\frac{k}{2}+p, \frac{1}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2+\lambda\right), \\ \sigma^{-2}|y,\theta,\mu,\upsilon^{-2} &\sim \text{Gamma}\left(\frac{n}{2}+q, \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2+\kappa\right), \\ \mu|y,\theta,\sigma^{-2},\upsilon^{-2} &\sim \text{N}\left(\frac{k\tau^2}{k\tau^2+\upsilon^2}\bar{\theta}+\frac{\upsilon^2}{k\tau^2+\upsilon^2}m, \frac{\tau^2\upsilon^2}{k\tau^2+\upsilon^2}\right), \\ \theta_i|y,\theta_{-i},\mu,\sigma^{-2},\upsilon^{-2} &\sim \text{N}\left(\frac{n\upsilon^2}{n\upsilon^2+\sigma^2}\bar{y}_i+\frac{\sigma^2}{n\upsilon^2+\sigma^2}\mu, \frac{\upsilon^2\sigma^2}{n\upsilon^2+\sigma^2}\right), \end{aligned}$$

gdzie, rzecz jasna, $n=\sum n_i$, $\bar{\theta}=\sum_i\theta_i/k$ i $\theta_{-i}=(\theta_k)_{k\neq i}$. Zwróćmy uwagę, że współrzędne wektora θ są warunkowo niezależne (pełny rozkład warunkowy θ_i nie zależy od θ_{-i}). Dzięki temy możemy w próbniku Gibbsa potraktować θ jako cały „blok" współrządnych i zmieniać „na raz".

Próbnik Gibbsa ma w tym modelu przestrzeń stanów $\mathcal{X}$ składającą się z punktów $x=(\theta,\mu,\sigma^{-2},\upsilon^{-2})\in\mathbb{R}^{k+3}$. Reguła przejścia próbnika w wersji systematycznej (duży krok „SystemPG"),

$$\underbrace{\left(\theta,\mu,\sigma^{-2},\upsilon^{-2}\right)}_{X_t} \longmapsto \underbrace{\left(\theta,\mu,\sigma^{-2},\upsilon^{-2}\right)}_{X_{t+1}},$$

jest złożona z następujących „małych kroków":

— Wylosuj $\upsilon^{-2} \sim p(\upsilon^{-2}|y, \theta, \mu, \sigma^{-2}) = \text{Gamma}(\ldots)$,
— Wylosuj $\sigma^{-2} \sim p(\sigma^{-2}|y, \theta, \mu, \upsilon^{-2}) = \text{Gamma}(\ldots)$,
— Wylosuj $\mu \sim p(\mu|y, \theta, \sigma^{-2}, \upsilon^{-2}) = \text{N}(\ldots)$,
— Wylosuj $\theta \sim p(\theta|y, \mu, \sigma^{-2}, \upsilon^{-2}) = \text{N}(\ldots)$.

Łańcuch Markowa jest zbieżny do rozkładu *a posteriori*:

$$X_t \to \pi(\cdot) = p(\theta, \mu, \sigma^{-2}, \upsilon^{-2}|y).$$

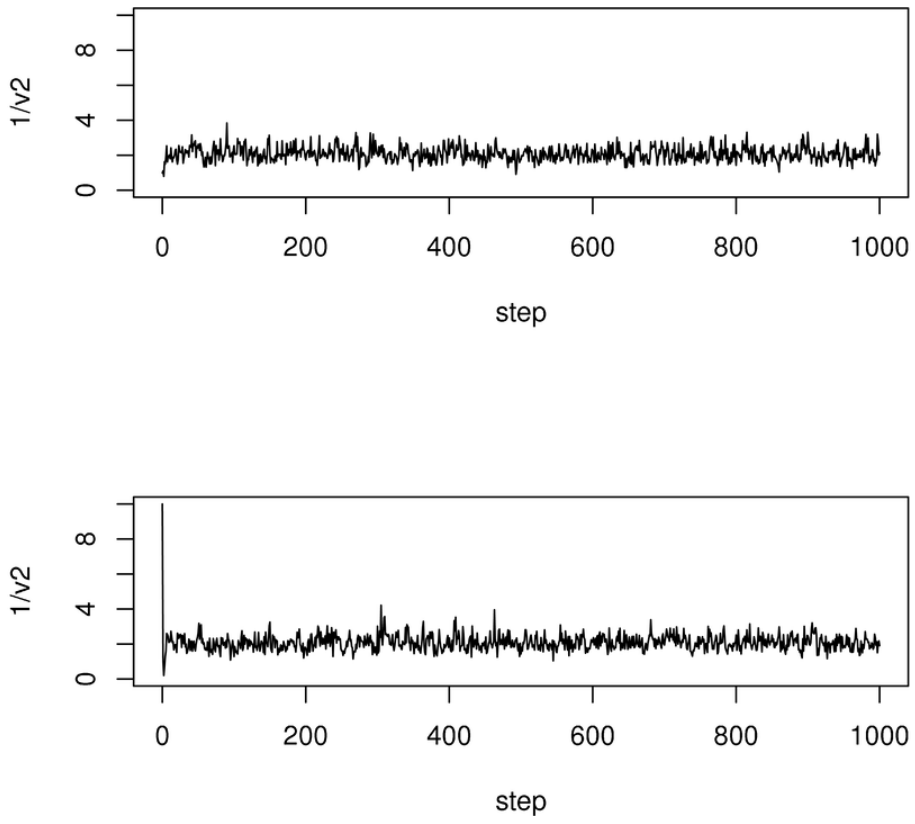
Najbardziej interesujące są w tym hierarchicznym modelu zmienne θ_i (pozostałe zmienne można uznać za „parametry zkłócające). Dla ustalenia uwagi zajmijmy się zmienną θ_1 (powiedzmy, wartością średnią w pierwszym małym obszarze). **Estymator bayesowski** jest to wartość oczekiwana *a posteriori* tej zmiennej:

$$\mathbb{E}(\theta_1|y) = \int \cdots \int \theta_1 p(\theta, \mu, \sigma^{-2}, \upsilon^{-2}|y) \mathrm{d}\theta_2 \cdots \mathrm{d}\theta_k \mathrm{d}\mu \mathrm{d}\sigma^{-2} \mathrm{d}\upsilon^{-2}.$$

Aproksymacją MCMC interesującej nas wielkości są średnie wzdłuż trajektorii łańcucha:

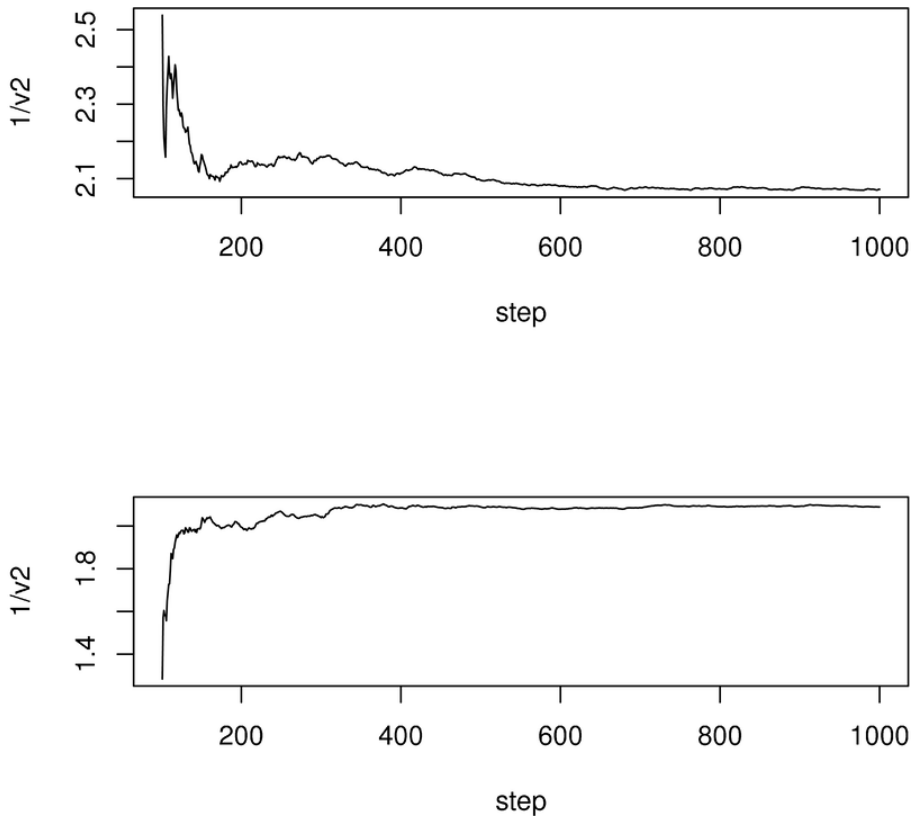
$$\theta_1(X_0), \theta_1(X_1), \ldots, \theta_1(X_t), \ldots$$

gdzie $\theta_1(x) = \theta_1$ dla $x = (\theta_1, \ldots, \theta_k, \mu, \sigma^{-2}, \upsilon^{-2})$.



Rysunek 12.1. Trajektorie zmiennej υ^{-2} dla dwóch punktów startowych.

Na Rysunku 12.1 pokazane są dwie przykładowe trajektorie współrzędnej υ^{-2} dla PG poruszającego się po przestrzeni $k + 3 = 1003$ wymiarowej (model uwzględniający 1000 małych obszarów). Dwie trajektorie odpowiadają dwum różnym punktom startowym. Dla innych zmiennych rysunki wyglądają bardzo podobnie. Uderzające jest to, jak szybko trajektoria zdaje się „osiągać" rozkład stacjonarny, przynajmniej wizualnie. Na Rysunku 12.2 pokazane są kolejne „skumulowane" średnie dla tych samych dwóch trajektorii zmiennej υ^{-2}.



Rysunek 12.2. Skumulowane średnie zmiennej v^{-2} dla dwóch punktów startowych.

12.2. Estymatory największej wiarogodności

Metody MCMC w połączeniu z ideą losowania istotnego (Rozdział 8) znajdują zastosowanie również w statystyce „częstościowej", czyli nie-bayesowskiej. Pokażę przykład, w którym oblicza się meodami Monte carlo estymator największej wiarogodności.

12.2.1. Model auto-logistyczny

Niech $x = (x_1, \ldots, x_d)$ będzie wektorem (konfiguracją) binarnych zmiennych losowych na przestrzeni $\mathcal{X} = \{0,1\}^d$. Rozważmy następujący *rozkład Gibbsa*:

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{i,j=1}^{d} \theta_{ij} x_i x_j \right\}.$$

Rolę parametru gra macierz (θ_{ij}). Zakłąd się, że jest to maciezrz symetryczna . Stała normująca $Z(\theta) = \sum_{x\in\mathcal{X}} \exp\left\{\sum_{i,j=1}^{d} \theta_{ij} x_i x_j\right\}$ jest typowo (dla dużych d) *niemożliwa do obliczenia*. Stanowi to, jak dalej zobaczymy, poważny problem dla statystyków.

Przykład 12.2 (Statystyka przestrzenna)**.** W zastosowaniach „przestrzennych" indeks $i \in \{1, \ldots, d\}$ interpretuje się jako „miejsce". Zbiór miejsc wyposażony jest w strukturę grafu. Krawędzie łączą miejsca „sąsiadujące". Piszemy $i \sim j$. Tego typu modele mogą opisywać na przykład rozprzestrzenianie się chorób lub występowanie pewnych gatunków. Wartośc $x_i = 1$ oznacza obecność gatunku lub występowanie choroby w miejscu i. Najprostszy model zakła-

da, że każda zmienna x_i zależy tylko od swoich „sąsiadów" i to w podobny sposób w całym rozpatrywanym obszarze. W takim modelu mamy tylko dwa parametry $\theta = (\theta_0, \theta_1)$:

$$\theta_{ij} = \begin{cases} 0 & i \not\sim j, i \neq j; \\ \theta_1 & i \sim j; \\ \theta_0 & i = j. \end{cases}$$

Parametr th_0 opisuje „skłonność" pojedynczej zmiennej do przyjmowania wartości 1, zaś parametr θ_1 odpowiada za zależność od zmiennych sąsiadujących (zakaźność choroby, powiedzmy). W typowej dla statystyki przestrzennej sytuacji, rozpatruje się nawet dziesiątki tysięcy „miejsc". Stała $Z(\theta)$ jest wtedy sumą niewyobrażalnie wielu (dokładnie 2^d) składników.

Próbnik Gibbsa w modelu auto-logistycznym

„Pełne" rozkłady warunkowe (*full conditionals*) są w modelu auto-logistycznym bardzo proste:

$$p_\theta(x_i = 1 \mid x_{-i}) = \frac{\exp\left(\theta_{ii} + \sum\limits_{\substack{j=1 \\ j\neq i}}^{d} \theta_{ij} x_j\right)}{1 + \exp\left(\theta_{ii} + \sum\limits_{\substack{j=1 \\ j\neq i}}^{d} \theta_{ij} x_j\right)},$$

gdzie $x_{-i} = (x_j, j \neq i)$. Zatem:

— Symulowanie $x \sim p_\theta$ jest łatwe za pomocą próbnika Gibbsa (PG):
$x_1 \sim p_\theta(x_1|x_{-1})$,
$x_2 \sim p_\theta(x_2|x_{-2})$, ...

Aproksymacja wiarogodności

Estymator największej wiarogodności obliczany metodą Monte Carlo został zaproponowany w pracy Geyer and Thopmpson (1992, *JRSS*). Rozważmy bardziej ogólną wykładniczą rodzinę rozkładów prawdopodobieństwa:

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp\left[\theta^\top T(x)\right],$$

gdzie $T(x)$ jest wektorem statystyk dostatecznych. Rozkłady autologistyczne tworzą rodzinę wykładniczą. Wystarczy θ ustawić w wektor, statystykami $T(x)$ są x_i i $x_i x_j$. Logarytm wiarogodności $L(\theta) = \log p_\theta(x)$ jest dany pozornie prostym wzorem:

$$L(\theta) = \theta^\top T(x) - \log Z(\theta).$$

Kłopot jest ze stałą normującą $Z(\theta)$, która w bardziej skomplikowanych modelach jest trudna lub wręcz niemożliwa do obliczenia. Wspomnieliśmy, że tak właśnie jest dla typowych modeli autologistycznych. Istnieje jednak prosty sposób aproksymacji stałej $Z(\theta)$ metodą MC. Istotnie, wybierzmy (w zasadzie dowolny) punkt θ_* i zauważmy, że

$$Z(\theta) = \sum_x \exp\left[\theta^\top T(x)\right] = \sum_x \exp\left[(\theta - \theta_*)^\top T(x)\right] \exp\left[\theta_*^\top T(x)\right].$$

Pozwala to wyrazić wielkość $Z(\theta)/Z(\theta_*)$ jako wartość oczekiwaną względem rozkładu p_{θ_*}:

$$\begin{aligned} Z(\theta)/Z(\theta_*) &= \sum_x \exp\left[(\theta - \theta_*)^\top T(x)\right] p_{\theta_*}(x) \\ &= \mathbb{E}_{\theta_*} \exp\left[(\theta - \theta_*)^\top T(x)\right]. \end{aligned}$$

Powyższe rozważania nie są niczym nowym: to jest po prostu przedstawiony w Rozdziale 8 schemat *losowania istotnego*, w specjalnym przypadku rodziny wykładniczej. Jeśli celem jest maksymalizacja wiarogodności $Z(\theta)$, to nieznana stała $Z(\theta_*)$ zupełnie nie przeszkadza (bo interesującą nas funkcję wiarygodności wystarczy znać z dokładnością do stałej).

Jeśli umiemy generować zmienne losowe o rozkładzie p_{θ_*}, to sposób przybliżonego obliczania $\hat{Z}(\theta)$ jest w zasadzie oczywisty. Generujemy próbkę MC: $x_*(k) \sim p_{\theta_*}$, $k = 1, \ldots, n_*$ gdzie θ_* jest *w zasadzie* dowolne, zaś n_* jest możliwie największe. Obliczamy

$$\hat{Z}(\theta)/Z(\theta_*) = \frac{1}{n_*} \sum_{k=1}^{n_*} \exp\left[(\theta - \theta_*)^\top T(x_*(k))\right].$$

Aproksymacja logarytmu wiarogodności polega na wstawieniu $\hat{Z}(\theta)$ w miejsce $\hat{Z}(\theta)$ we wzorze na $L(\theta) = \log p_\theta(x)$:

$$\hat{L}_{\mathrm{MC}}(\theta) = \theta^\top T(x) - \log \underbrace{\sum_{k=1}^{n_*} \exp[(\theta - \theta_*)^\top T(x_*(k))]}_{\text{przybliżenie MC } Z(\theta)} + \text{const.}$$

Pozostaje wiele szczegółów do dopracowania. Umiemy obliczać wiarogodność jako funkcję θ, ale trzeba jeszcze znaleźć maksimum tej funkcji. Dobór θ_* nie wpływa na poprawność rozumowania ale może mieć zasadniczy wpływ na efektywność algorytmu. Nie będziemy się w to zagłębiać. Wspomnimy tylko, jak omawiana metoda jest związana z *markowowskimi* metodami Monte Carlo, MCMC. Wróćmy do rozkładu auto-logistycznego. Jak losować próbkę z tego rozkładu? Tu przychodzi z pomocą próbnik Gibbsa. W rzeczy samej, mamy do czynienia z kolejną aproksymacją: PG jest tylko *przybliżonym* algorytmem generowania z rozkładu p_{θ_*}, które to losowanie pozwala obliczać wiarogodność *w przybliżeniu*. Momo wszystko, metody MCMC oferują pewne wyjście lepsze niż bezradne rozłożenie rąk.

13. Markowowskie Monte Carlo IV. Pola losowe

13.1. Definicje

Niech $(\mathcal{S}, \mathcal{E})$ będzie nieskierowanym grafem. Wyobraźmy sobie, że elementy $s \in S$ reprezentują „miejsca" w przestrzeni lub na płaszczyźnie, zaś krawędzie grafu łączą miejsca „sąsiadujące" ze sobą. Taka interpretacja jest związana z zastosowaniami do statystyki „przestrzennej" i przetwarzania obrazów. Model, który przedstawimy ma również zupełnie inne interpretacje, ale pozostaniemy przy sugestywnej terminologii „przestrzennej":
— $\mathcal{S}$ — zbiór miejsc,
— $\{s, t\} \in \mathcal{E}$ — miejsca s i t sąsiadują — będziemy wtedy pisać $s \sim t$,
— $\partial t = \{s : s \sim t\}$ — zbiór sąsiadów miejsca t.
Niech $\Lambda = \{1, \ldots, l\}$ będzie skończonym zbiorem. Powiedzmy, że elementy $a \in \Lambda$ są „kolorami" które mogą być przypisane elementom zbioru $\mathcal{S}$. Konfiguracją nazywamy dowolną funkcję $x : \mathcal{S} \to \Lambda$. Będziemy mówić że $x_s = x(s)$ jest s-tą współrzędną konfiguracji x i stosować oznaczenia podobne jak dla wektorów:

$$x = (x_s) = (x_s : s \in \mathcal{S}).$$

W zadaniach przetwarzania obrazów, miejsca są pikslami na ekranie i konfigurację utożsamiamy z ich pokolorowaniem, a więc z cyfrową reprezentacją obrazu. Zbiór Λ gra rolę „palety kolorów". Niekiedy założenie o skończoności zbioru Λ staje się niewygodne. Dla czarno-szaro-białych obrazów „pomalowanych" różnymi odcieniami szarości, wygodnie przyjąć, że $\Lambda = [0, 1]$ lub $\lambda = [0, \infty[$. Tego typu modyfikacje są dość oczywiste i nie będę się nad tym zatrzymywał. Dla ustalenia uwagi, wzory w tym podrozdziale dotyczą przypadku skończonego zbioru „kolorów". Przestrzenią konfiguracji jest zbiór $\mathcal{X} = \Lambda^{\mathcal{S}}$. Dla konfiguracji x i miejsca t, niech
— $x_{-t} = (x_s : s \neq t) = (x_s : s \in \mathcal{S} \setminus \{t\})$ — konfiguracja z pominiętą t-tą współrzędną,
— $x_{\partial t} = (x_s : s \in \partial t)$ — konfiguracja ograniczona do sąsiadów miejsca t.

Jeśli $H : \mathcal{X} \to \mathbb{R}$ i $\beta \geqslant 0$ to **rozkładem Gibbsa** nazywamy rozkład prawdopodobieństwa na przestrzeni konfiguracji dany wzorem

$$\pi_\beta(x) = \frac{1}{Z(\beta)} \exp[-\beta H(x)].$$

Ze względu na inspiracje pochodzące z fizyki statystycznej, funkcję H nazywamy energią, β jest (z dokładnością do stałej) odwrotnością temperatury. Stała normująca wyraża się wzorem

$$Z(\beta) = \sum_{x \in \mathcal{X}} \exp[-\beta H(x)].$$

i jest typowo *niemożliwa do obliczenia.*

Oczywiście, każdy rozkład prawdopodobieństwa π na $\mathcal{X}$ dale się zapisać jako rozkład Gibbsa, jeśli położyć $H(x) = -\log \pi(x)$, $\beta = 1$ i umownie przyjąć, że $-\log 0 = \infty$ (czyli konfiguracje niemożliwe mają nieskończoną energię). Nie o to jednak chodzi. Ciekawe są rozkłady Gibbsa, dla których fukcja energii ma specjalną postać związaną z topologią grafu „sąsiedztw". Ograniczymy

się do ważnej podklasy markowowskich pól losowych (MPL), mianowicie do sytuacji gdy energia jest sumą „oddziaływań" lub „interakcji" między parami miejsc sąsiadujących i składników zależnych od pojedynczych miejsc. Dokładniej, załóżymy że

$$H(x) = \sum_{s\sim t} V(x_s, x_t) + \sum_s U_s(x_s), \tag{13.1}$$

dla pewnych funkcji $V : \Lambda\times\Lambda \to \mathbb{R}$ i $U_s : \Lambda \to \mathbb{R}$. Funkcja $V(x_s, x_t)$ opisuje „potencjał interakcji pomiędzy s i t", zaś $U_s(a)$ jest wielkością związaną z „tendencją miejsca s do przybrania koloru a. Zwróćmy uwagę, że potencjał V jest jednorodny ($V(a, b)$ zależy tylko od „kolorów" $a, b \in \Lambda$ ale nie od miejsc), zaś $U_s(a)$ może zależeć zarówno od $a \in \Lambda$ jak i od $s \in \mathcal{S}$. W modelach fizyki statystycznej zazwyczaj $U_s(a) = U(a)$ jest jednorodnym „oddziaływaniem zewnętrznym" ale w modelach rekonstrukcji obrazów nie można tego zakładać.

Przykład 13.1 (Model Pottsa)**.** Niech Λ będzie zbiorem skończonym i

$$H(x) = J\sum_{s\sim t} \mathbb{I}(x_s \neq x_t).$$

Ta funkcja opisuje „tendencję sąsiednich miejsc do przybierania tego samego koloru". Jeśli $J > 0$ to preferowane są konfiguracje złożone z dużych, jednobarwnych plam.

13.2. Generowanie markowowskich pól losowych

Użyteczność MPL w różnorodnych zastosowaniach związana jest z istnieniem efektywnych algorytmów symulacyjnych. Wszystko opiera się na próbniku Gibbsa i następującym prostym fakcie.

Twierdzenie 13.1 (Pełne rozkłady warunkowe dla MPL)**.** *Jeżeli π_β jest rozkładem Gibbsa z energią daną wzorem (13.1), to*

$$\pi_\beta(x_s|x_{-s}) = \pi_\beta(x_s|x_{\partial s}) = \frac{1}{Z_s(\beta)} \exp\left[-\beta H_s(x)\right],$$

gdzie

$$H_s(x) = \sum_{t\in\partial s} V(x_s, x_t) + U(x_s),$$

$Z_s(\beta) = \sum_{a\in\Lambda} \exp[H_s(x_{a\rightsquigarrow s})$. *Symbol $x_{a\rightsquigarrow s}$ oznacza konfigurację powstałą z x przez wpisanie koloru a w miejscu s.*

Dowód. Skorzystamy z elementarnej definicji prawdopodobieństwa warunkowego (poniżej piszemy $\pi_\beta(\cdot) = \pi(\cdot)$, bo parametr β jest ustalony):

$$\begin{aligned}
\pi(x_s|x_{-s}) &= \frac{\pi(x)}{\pi(x_{-s})} = \frac{\pi(x)}{\sum\limits_a \pi(x_{a\rightsquigarrow s})} \\
&= \frac{\exp -\beta H(x)}{\sum_a \exp -\beta H(x_{a\rightsquigarrow s})} \\
&= \frac{\exp -\beta \left(\sum\limits_{t:t\sim s} V(x_s, x_t) + \sum\limits_{t\sim w, t\neq s, w\neq s} V(x_t, x_w) + U_s(x_s) + \sum\limits_{t\neq s} U_t(x_t) \right)}{\sum\limits_a \exp -\beta \left(\sum\limits_{t:t\sim s} V(a, x_t) + \sum\limits_{t\sim w, t\neq s, w\neq s} V(x_t, x_w) + U_s(a) + \sum\limits_{t\neq s} U_t(x_t) \right)} \\
&= \frac{\exp -\beta \left(\sum\limits_{t:t\sim s} V(x_s, x_t) + U_s(x_s) \right)}{\sum\limits_a \exp -\beta \left(\sum\limits_{t:t\sim s} V(a, x_t) + U_s(a) \right)} \\
&= \frac{\exp -\beta H_s(x)}{Z_s(\beta)}.
\end{aligned}$$

Ponieważ otrzymany wynik zależy tylko od x_s i $x_{\partial s}$, więc $\pi(x_s|x_{-s}) = \pi(x_s|x_{\partial s})$. Ten wniosek jest pewną formą własności Markowa. □

Zauważmy, że obliczenie $H_s(x)$ jest łatwe, bo suma $\sum_{t\in\partial s} \cdots$ zawiera tylko tyle składników, ile jest sąsiadów miejsca s. Obliczenie $Z_s(\beta)$ też jest łatwe, bo suma $\sum_{a\in\Lambda} \cdots$ zawiera tylko $l = |\Lambda|$ składników. Ale nawet nie musimy obliczać stałej normującej $Z_s(\beta)$ żeby generować z rozkładu

$$\pi(x_s = a|x_{\partial s}) \propto exp - \beta \left(\sum_{t:t\sim s} V(a, x_t) + U_s(a) \right). \tag{13.2}$$

Na tym opiera się implementacja próbnika Gibbsa. Wersję PG z „systemstycznym przeglądem miejsc" można zapisać tak:

Listing.

for $s \in \mathcal{S}$ **do**
begin
Gen $a \sim \pi(x_s = \cdot|x_{\partial s})$;
$x := x_{a\rightsquigarrow s}$
end

Faktycznie już ten algorytm spotkaliśmy w Podrozdziale 12.2, dla szczególnego przypadku modelu auto-logistycznego.

13.3. Rekonstrukcja obrazów

Bayesowski model rekonstrukcji obrazów został zaproponowany w pracy Gemana i Gemana w 1987 roku. Potem zdobył dużą popularność i odniósł wiele sukcesów. Model łączy idee zaczerpnięte ze statystyki bayesowskiej i fizyki statystycznej. Cyfrową reprezentację obrazu utożsamiamy z konfiguracją kolorów na wierzchołkach grafu, czyli z elementem przestrzeni $\mathcal{X} = \Lambda^{\mathcal{S}}$, zdefiniowanej w Podrozdziale 13.1. Przyjmijmy, że „idealny obraz", czyli to co chcielibyśmy zrekonstruować jest konfiguracją $x \in \mathcal{X}$. Niestety, obraz jest „zakłócony" lub „zaszumiony".

Możemy tylko obserwować konfigurację y reprezentującą zakłócony obraz. Zbiór kolorów w obrazie y nie musi być identyczny jak w obrazie x. Powiedzmy, że $y \in \mathcal{Y} = \Gamma^{\mathcal{S}}$. Ważne jest to, że zniekształcenie modelujemy probabilistycznie przy pomocy rodziny rozkładów warunkowych $f(y|x)$. Dodatkowo zakładamy, że obraz x pojawia się losowo, zgodnie z rozkładem prawdopodobieństwa $\pi(x)$. Innymi słowy, „idealny" obraz x oraz „zniekształcony" obraz y traktujemy jako realizacje zmiennych losowych $X : \Omega \to \mathcal{X}$ i $Y : \Omega \to \mathcal{Y}$,

$$\pi(x) = \mathbb{P}(X = x), \qquad f(y|x) = \mathbb{P}(Y = y|X = x).$$

W ten sposób buduje się statystyczny model bayesowski, w którym
— Y jest obserwowaną zmienną losową,
— x jest nieznanym parametrem traktowanym jako zmienna losowa X.
Oczywiście, π gra rolę rozkładu *a priori*, zaś f jest wiarogodnością. Być może użycie literki x na oznaczenie parametru jest niezgodne z tradycyjnymi oznaczeniami statystycznymi, ale z drugiej strony jest wygodne. Wzór Bayesa mówi, że rozkład *a posteriori* jest następujący.

$$\pi_y(x) = \mathbb{P}(X = x|Y = y) \propto f(y|x)\pi(x).$$

Pomysł Gemana i Gemana polegał na tym, żeby modelować rozkład *a priori* π jako MPL. Załóżmy, że π jest rozkładem Gibbsa,

$$\pi(x) \propto \exp\left(-H(x)\right), \tag{13.3}$$

gdzie

$$H(x) = J \sum_{s \sim t} V(x_s, x_t). \tag{13.4}$$

Energia „*a priori*" zawiera tu tylko składniki reprezentujące oddziaływania między parami miejsc sąsiednich. Funkcja $V(a, b)$ zazwyczaj ma najmniejszą wartość dla $a = b$ i rośnie wraz z „odległością" między a i b (jakkolwiek tę odległość zdefiniujemy). W ten sposób „nagradza" konfiguracje w których sąsiednie miejsca są podobnie pokolorowane. Im większy parametr $J > 0$, tym bardziej prawdopodobne są obrazy zawierające jednolite plamy kolorów.

Trzeba jeszcze założyć coś o „wiarogodności" f. Dla uproszczenia opiszę tylko najprostszy model, w którym kolor y_s na obserwowanym obrazie zależy tylko od koloru x_s na obrazie idealnym. Intuicyjnie znaczy to, że „zaszumienie" ma ściśle lokalny charakter. Matematycznie znaczy to, że

$$f(y|x) = \prod_s f(y_s|x_s)$$

(pozwolę sobie na odrobinę nieścisłości aby uniknąć nowego symbolu na oznaczenie $f(y_s|x_s)$). Zapiszemy teraz „wiarogodność" f w postaci zlogarytmowanej. Jeśli położymy $-\log f(y_s|x_s) = U_s(x_s)$ pamiętając, że y jest w świecie bayesowskim ustalone, to otrzymujemy następujący wzór:

$$\pi_y(x) \propto \exp\left(-H_y(x)\right), \tag{13.5}$$

gdzie

$$H_y(x) = J \sum_{s \sim t} V(x_s, x_t) - \sum_s \log f(y_s|x_s). \tag{13.6}$$

Okazuje się zatem, że rozkład *a posteriori* ma podobną postać do rozkładu *a priori*. Też jest rozkładem Gibbsa, a różnica polega tylko na dodaniu składników reprezentujących oddziaływania zewnętrzne $U_s(x_s) = -\log f(y_s|x_s)$. Pamiętajmy przy tym, że y jest w świecie bayesowskim ustalone. W modelu rekonstrukcji obrazów „oddziaływania zewnętrzne" zależą od y i „wymuszają podobieństwo" rekonstruowanego obrazu do obserwacji. Z kolei „oddziaływania między parami" są odpowiedzialne za wygładzenie obrazu. Lepiej to wyjaśnimy na przykładzie.

Przykład 13.2 (Losowe „przekłamanie koloru" i wygładzanie Pottsa)**.** Załóżmy, że $\Lambda = \{1, \ldots, l\}$ jest naprawdę paletą kolorów, na przykład

$$\Lambda = \{Czerwony, Niebieski, Pomaraczowy, Zielony\}.$$

Przypuśćmy, że mechanizm losowego „przekłamania" polega na tym, że w każdym pikslu, kolor obecny w idealnym obrazie x jest z prawdopodobieństwem $1 - \varepsilon$ niezmieniony, a z prawdopodobieństwem ε zmienia się na losowo wybrany inny kolor. Tak więc zarówno x jak i y należą do tej samej przestrzeni $\Lambda^{\mathcal{S}}$,

$$f(y_s|x_s) = \begin{cases} 1-\varepsilon & \text{dla } y_s = x_s; \\ \varepsilon/(l-1) & \text{dla } y_s \neq x_s. \end{cases}$$

Można za rozkład *a priori* przyjąć rozkład Pottsa z Przykładu 13.1. Rozkład *a posteriori* ma funkcję energii daną następującym wzorem (z $J > 0$):

$$H_y(x) = J \sum_{s \sim t} \mathbb{I}(x_s \neq x_t) - \sum_s \left[\log(1-\varepsilon)\mathbb{I}(x_s = y_s) + \log(\varepsilon/(l-1))\mathbb{I}(x_s \neq y_s)\right].$$

Pierwszy składnik w tym wzorze pochodzi od rozkładu a priori (z modelu Pottsa) i „nagradza" konfiguracje w których dużo sąsiednich punktów jest pomalowanych na ten sam kolor. Powoduje to, że obrazy x składające się z jednolotych dużych „plam" są preferowane. Drugi składnik pochodzi od obserwowanegj konfiguracji y i jest najmniejszy dla $x = y$. Powoduje to, ze obrazy x mało się różniące od y są bardziej prawdopodobne. Rozkład *a posteriori* jest pewnym kompromisem pomiędzy tymi dwoma konkurującymi składnikami. Parametr J jest „wagą" pierwszego składnika i dlatego odgrywa rolę „parametru wygładzającego". Im większe J tym odtwarzany obraz będzie bardziej regularny (a tym mnie będzie starał się upodobnić do y). I odwrotnie, małe J powoduje ściślejsze dopasowanie x do y ale mniejszą „regularność" x.

Jeszcze lepiej to samo widać na przykładzie tak zwanego „szumu gaussowskiego".

Przykład 13.3 (Addytywny szum gaussowski)**.** Załóżmy, że x jest konfiguracją „poziomów szarości" czyli, powiedzmy, $\Lambda \subseteq [0, \infty[$. Mechanizm losowego „zaszumienia" polega na tym, że zamiast poziomu szarości x_s obserwujemy $y_s \sim \mathrm{N}(x_s, \sigma^2)$. Innymi słowy,

$$f(y_s|x_s) \propto \exp\left[-\frac{1}{2\sigma^2}(y_s - x_s)^2\right].$$

Przestrzenią obserwaowanych konfiguracji y jest tutaj (formalnie) $\mathbb{R}^{\mathcal{S}}$ (faktycznie, raczej $[0, \infty[^{\mathcal{S}}$). Rozkład *a posteriori* ma funkcję energii daną następującym wzorem:

$$H_y(x) = J \sum_{s \sim t} V(x_s \neq x_t) + \frac{1}{2\sigma^2} \sum_s (y_s - x_s)^2.$$

Jeśli rozpatrujemy model ze skończoną liczbą poziomów szarości dla konfiguracji x to można pierwszy składnik określić tak jak w poprzednim przykładzie, czyli zapożyczyć z modelu Pottsa. Bardziej naturalne jest określenie $V(a, b)$ w taki sposób, aby większe różnice pomiędzy poziomami a i b były silniej karane. parametr J jest, jak poprzednio, odpowiedzialny zaa stopień wygładzenia.

14. Markowowskie Monte Carlo V. Elementarna teoria łańcuchów Markowa

14.1. Podstawowe określenia i oznaczenia

W tym rozdziale rozważamy jednorodny łańcuch Markowa $X_0, X_1, \ldots, X_n, \ldots$ na skończonej przestrzeni stanów $\mathcal{X} = \{1, \ldots, d\}$. Będziemy posługiwać się wygodną i zwięzłą notacją wektorowo-macierzową. Macierz przejścia o wymiarach $d \times d$ oznaczamy $P = (P(x,y))_{x,y\in\mathcal{X}}$. Rozkład początkowy utożsamiamy z wektorem wierszowym $\xi^\top = (\xi(1), \ldots, \xi(x), \ldots, \xi(d))$. W dalszym ciągu, mówiąc o łańcuchu Markowa, będziemy mieli na myśli ustaloną macierz przejścia P i dowolnie wybrany rozkład początkowy ξ. Przyjmujemy oznaczenie $\mathbb{P}_\xi(\cdot)$. W szczególności, $\mathbb{P}_x(\cdot) = \mathbb{P}(\cdot|X_0 = x)$, dla $x \in \mathcal{X}$. Analogicznie będziemy oznaczali wartość oczekiwaną: $\mathbb{E}_\xi$ lub $\mathbb{E}_x$. Zauważmy, że $\mathbb{P}(X_{n+2} = y|X_n = x) = \sum_z P(x,z)P(z,y) = P^2(x,y)$. Ogólniej, macierz przejścia w m krokach jest m-tą potęgą macierzy P:

$$\mathbb{P}(X_{n+m} = y|X_n = x) = P^m(x,y).$$

Rozkład *brzegowy* zmiennej losowej X_n jest wektorem $\xi^\top P^n$:

$$\mathbb{P}(X_n = x) = (\xi^\top P^n)(x).$$

Z macierzą stochastyczną P związany jest graf skierowany opisujący możliwe przejścia łańcucha. Jest to graf $(\mathcal{X}, \mathcal{E})$, gdzie zbiorem wierzchołków jest przestrzeń stanów, zaś $\mathcal{E} \subseteq \mathcal{X}\times\mathcal{X}$ jest zbiorem takich par (x,y), że $P(x,y) > 0$. Pojęcia, o których teraz będziemy mówić, są związane tylko ze strukturą grafu, czyli z położeniem niezerowych elementów macierzy przejścia.

Mówimy, że stan y jest *osiągalny* ze stanu x, jeśli $x = y$ lub istnieje liczba naturalna $n \geqslant 1$ i ciąg stanów $x = x_0, x_1, x_2, \ldots, x_{n-1}, x_n = y$ taki, że $P(x_{i-1}, x_i) > 0$ dla $i = 1, \ldots, n$. Równoważnie, $P^n(x,y) > 0$ dla pewnego $n \geqslant 0$. Będziemy stosowali oznaczenie „$x \to y$”.

Stany x i y *komunikują* się, co zapiszemy „$x \leftrightarrow y$”, jeśli $x \to y$ i $y \to x$.

Stan x jest *istotny*, jeśli dla każdego y takiego, że $x \to y$ mamy $y \to x$. W przeciwnym razie stan x nazywamy *nieistotnym*. Stan x jest więc nieistotny, jeśli istnieje stan y do którego można przejść z x, ale nie można wrócić do x. Zbiór stanów istotnych oznaczymy przez $\mathcal{C}$, zaś zbiór stanów nieistotnych przez $\mathcal{T}$. Zbiór $\mathcal{C}$ jest (dla łańcucha o skończonej przestrzeni stanów) zawsze niepusty. Zbiór $\mathcal{T}$ może być pusty.

Jeśli $x \leftrightarrow y$ dla dowolnych $x, y \in \mathcal{X}$, czyli wszystkie stany komunikują się, to łańcuch nazywamy *nieprzywiedlnym*. Oczywiście, wszystkie stany łańcucha nieprzywiedlnego są istotne, $\mathcal{X} = \mathcal{C}$. Lańcuch nieprzywiedlny nazywamy *nieokresowym*, jeśli dla dowolnych $x, y \in \mathcal{X}$ istnieje n_0 takie, że dla każdego $n \geqslant n_0$ mamy $P^n(x,y)$. Wspomnijmy, że przyjęta przez nas definicja nieokresowości bywa formułowana w nieco inny, ale równoważny sposób. Dla naszych potrzeb wystarczy następujące proste spostrzeżenie: jeśli łańcuch jest nieprzywiedlny i $P(x,x) > 0$ dla przynajmniej jednego stanu x, to łańcuch jest nieokresowy. Latwo zauważyć, że dla łańcucha nieprzywiedlnego i nieokresowego, dla dostatecznie dużych n macierz P^n ma wszystkie elementy niezerowe. Wynika to z faktu, że rozważamy łańcuchy ze skończoną przestrzenią stanów.

Jeśli $x \leftrightarrow y$ dla dowolnych $x, y \in \mathcal{C}$, czyli wszystkie stany *istotne* komunikują się, to łańcuch nazywamy *jednoklasowym*. Lańcuch jednoklasowy może mieć niepusty zbiór stanów nieistotnych $\mathcal{T}$. Lańcuch jednoklasowy możemy „przerobić" na nieprzywiedlny jeśli ograniczymy przestrzeń do stanów istotnych. Macierz $P_{|\mathcal{C}} = (P(x,y)_{x,y\in\mathcal{C}})$ jest, jak łatwo widzieć, stochastyczna.

Interesują nas głównie łańcuchy, które „zmierzają w kierunku położenia równowagi". Aby uściślić co to znaczy „równowaga", przypomnijmy pojęcie stacjonarności. Rozkład π jest stacjonarny jeśli dla każdego stanu y,

$$\pi(y) = \sum_x \pi(x)P(x,y).$$

W notacji macierzowej: $\pi^\top = \pi^\top P$. Stąd oczywiście wynika, że $\pi^\top = \pi^\top P^n$.

Poniższy prosty fakt można uzasadnić na wiele sposobów. W następnym podrozdziale przytoczymy, wraz z dowodem, piękne twierdzenie Kaca (Twierdzenie 14.2), które implikuje Twierdzenie 14.1.

Twierdzenie 14.1. *Jeśli łańcuch Markowa jest nieprzywiedlny, to istnieje dokładnie jeden rozkład stacjonarny π, przy tym $\pi(x) > 0$ dla każdego $x \in \mathcal{X}$.*

Z Twierdzenia 14.1 łatwo wynika następujący wniosek.

Wniosek 14.1. *Jeśli łańcuch Markowa jest jednoklasowy, to istnieje dokładnie jeden rozkład stacjonarny π, przy tym $\pi(x) > 0$ dla każdego $x \in \mathcal{C}$, czyli dla wszystkich stanów istotnych oraz $\pi(x) = 0$ dla każdego $x \in \mathcal{T}$, czyli dla stanów nieistotnych.*

Uwaga 14.1. Podkreślmy stale obowiązujące w tym rozdziale założenie, że *przestrzeń stanów jest skończona*. To założenie jest istotne w Twierdzeniu 14.1 i to samo dotyczy dalszych rozważań. Istnieją co prawda odpowiedniki sformułowanych tu twierdzeń dla przypadku ogólnej przestrzeni stanów (nieskończonej, a nawet „ciągłej" takiej jak $\mathbb{R}^d$) ale wymagają one dodatkowych, niełatwych do sprawdzenia założeń. Przystępny i bardzo elegancki wykład teorii łańcuchów Markowa na ogólnej przestrzeni stanów można znaleźć w pracy Nummelina [17]. Przeglądowy artykuł Robertsa i Rosenthala [20] zawiera dużo dodatkowych informacji na ten temat. Obie cytowane prace koncentrują się na tych własnościach łańcuchów, które są istotne z punktu widzenia algorytmów Monte Carlo. Z kolei piękna książka Brémaud [4] ogranicza się do przestrzeni dyskretnych (skończonych lub przeliczalnych).

14.2. Regeneracja

Przedstawimy w tym podrozdziale konstrukcję, która prowadzi do łatwch i eleganckich dowodów twierdzeń granicznych. Podstawowa idea jest następująca. Wyróżnia się jeden ustalony stan, powiedzmy $z \in \mathcal{X}$. W każdym momencie wpadnięcia w z, następuje „odnowienie" i dalsza ewolucja łańcucha jest niezależna od przeszłości.

Niech, dla $z \in \mathcal{X}$,

$$T^z = \min\{n > 0 : X_n = z\}. \tag{14.1}$$

Przyjmujemy przy tym naturalną konwencję: $T^z = \infty$, jeśli $X_n \neq z$ dla każdego $n \geqslant 1$. Zmienna losowa T^z jest więc czasem pierwszego dojścia do stanu z. Jeśli założymy, że łańcuch startuje z punktu z, to T^z jest czasem pierwszego powrotu.

Lemat 14.1. *Jeżeli łańcuch jest nieprzywiedlny, to istnieją stałe c i $\gamma < 1$ takie, że dla dowolnego rozkładu początkowego ξ,*

$$\mathbb{P}_\xi(T^z > n) \leqslant c\gamma^n.$$

Dowód. Dla uproszczenia przyjmijmy dodatkowe założenie, że łańcuch jest nieokresowy. Wtedy dla dostatecznie dużych k wszystkie elementy macierzy P^k są niezerowe. Ustalmy k i znajdźmy liczbę $\delta > 0$ taką, że $P^k(x,z) \geqslant \delta$ dla wszystkich x (jest to możliwe, bo łańcuch ma skończoną liczbę stanów). Dla dowolnego n, dobierzmy takie m, że $mk \leqslant n < (m+1)k$. Mamy wówczas

$$\begin{aligned}\mathbb{P}_\xi(T^y > n) &\leqslant \mathbb{P}_\xi(T^y > mk)\\ &\leqslant \mathbb{P}_\xi(X_0 \neq z, X_k \neq z, \ldots, X_{mk} \neq z)\\ &= \sum_{x_0\neq z, x_1 \neq z, \ldots, x_m \neq z} \xi(x_0)P^k(x_0,x_1)\cdots P^k(x_{m-1},x_m)\\ &\leqslant (1-\delta)^m \leqslant c\gamma^n,\end{aligned}$$

dla $\gamma = (1-\delta)^{1/k}$ i $c = (1-\delta)^{-1}$.

W przypadku łańcucha okresowego dowód nieco się komplikuje i, choć nie jest trudny, zostanie pominięty. □

Wniosek 14.2. *Dla łańcucha nieprzywiedlnego mamy $\mathbb{P}_\xi(T^z < \infty) = 1$, co więcej $\mathbb{E}_\xi T^z < \infty$, co więcej $\mathbb{E}_\xi (T^z)^k < \infty$ dla dowolnego k, a nawet $\mathbb{E}_\xi \exp(\lambda T^z) < \infty$ przynajmniej dla pewnych dostatecznie małych wartości $\lambda > 0$ (w istocie dla $\lambda < -\log\gamma$).*

Podamy teraz bardzo ciekawą interpretację rozkładu stacjonarnego, wykazując przy okazji jego istnienie (Twierdzenie 14.1). Ustalmy dowolnie wybrany stan z. Udowodnimy, że średni czas, spędzony przez łańcuch w stanie y pomiędzy wyjściem z z i pierwszym powrotem do z jest *proporcjonalny* do $\pi(y)$, prawdopodobieństwa stacjonarnego.

Twierdzenie 14.2 (Kaca)**.** *Załóżmy, że łańcuch jest nieprzywiedlny. Ustalmy $z \in \mathcal{X}$ i zdefiniujmy miarę α wzorem*

$$\alpha(y) = \mathbb{E}_z \sum_{i=0}^{T^z-1} \mathbb{I}(X_i = y) = \mathbb{E}_z \sum_{i=1}^{T^z} \mathbb{I}(X_i = y).$$

Wtedy:

(i) Miara α jest stacjonarna, czyli $\alpha^\top P = \alpha$.

(ii) Miara α jest skończona, $\alpha(\mathcal{X}) = \mathbb{E}_z(T^z) = m < \infty$.

(iii) Unormowana miara $\alpha/m = \pi$ jest jedynym rozkładem stacjonarnym.

Dowód. Dla uproszczenia będziemy pisali $T^z = T$, $\mathbb{P}_z = \mathbb{P}$ i $\mathbb{E}_z = \mathbb{E}$. Zauważmy, że

$$\begin{aligned}\alpha(x) &= \mathbb{E}\sum_{i=0}^{T-1} \mathbb{I}(X_i = x) = \mathbb{E}\sum_{i=0}^{\infty} \mathbb{I}(X_i = x, T > i)\\ &= \sum_{i=0}^{\infty} \mathbb{P}(X_i = x, T > i).\end{aligned}$$

Udowodnimy teraz (i). Jeśli $y \neq z$, to

$$\begin{aligned}\sum_x \alpha(x)P(x,y) &= \sum_x \sum_{i=0}^{\infty} \mathbb{P}(X_i = x, T > i)P(x,y) \\ &= \sum_{i=0}^{\infty} \sum_x \mathbb{P}(X_i = x, T > i)P(x,y) \\ &= \sum_{i=0}^{\infty} \mathbb{P}(X_{i+1} = y, T > i+1) = \sum_{i=1}^{\infty} \mathbb{P}(X_i = y, T > i) \\ &= \alpha(y),\end{aligned}$$

ponieważ $\mathbb{P}(X_0 = y) = 0$, bo $\mathbb{P}(X_0 = z) = 1$. Dla $y = z$ mamy z kolei

$$\begin{aligned}\sum_x \alpha(x)P(x,z) &= \sum_x \sum_{i=0}^{\infty} \mathbb{P}(X_i = x, T > i)P(x,z) \\ &= \sum_{i=0}^{\infty} \sum_x \mathbb{P}(X_i = x, T > i)P(x,z) \\ &= \sum_{i=0}^{\infty} \mathbb{P}(X_{i+1} = z, T = i+1) = \sum_{i=1}^{\infty} \mathbb{P}(T = i) \\ &= 1 = \alpha(z),\end{aligned}$$

co kończy dowód (i).

Część (ii) jest łatwa. Równość

$$\alpha(\mathcal{X}) = \sum_y \alpha(y) = \mathbb{E}T$$

wynika wprost z definicji miary α. Fakt, że $m = \mathbb{E}T < \infty$ jest wnioskiem z Lematu 14.1.

Punkt (iii): istnienie rozkładu stacjonarnego

$$\pi(y) = \frac{\alpha(y)}{m}.$$

jest natychmiastowym wnioskiem z (i) i (ii). Jednoznaczność rozkładu stacjonarnego dla jest nietrudna do bezpośredniego udowodnienia. Pozostawiamy to jako ćwiczenie. W najbardziej interesującym nas przypadku łańcucha nieokresowego, jednoznaczność wyniknie też ze Słabego Twierdzenia Ergodycznego, które udowodnimy w następnym podrozdziale. □

Odnotujmy ważny wniosek wynikający z powyższego twierdzenia:

$$\pi(z) = \frac{1}{\mathbb{E}_z(T^z)}.$$

Zjawisko odnowienia, czyli regeneracji pozwala sprowadzić badanie łańcuchów Markowa do rozpatrywania niezależnych zmiennych losowych, a więc do bardzo prostej i dobrze znanej sytuacji. Aby wyjaśnić to bliżej, zauważmy następującą oczywistą równość. Na mocy własności Markowa i jednorodności,

$$\begin{aligned}&\mathbb{P}(X_{n+1} = x_1, \ldots, X_{n+k} = x_k | T^z = n) \\ &\mathbb{P}(X_{n+1} = x_1, \ldots, X_{n+k} = x_k | X_n = z) \\ &\quad = \mathbb{P}_z(X_1 = x_1, \ldots, X_k = x_k).\end{aligned}$$

Zatem warunkowo, dla $T^z = n$, łańcuch „regeneruje się w momencie n" i zaczyna się zachowywać dokładnie tak, jak łańcuch który wystartował z punktu z w chwili 0. Niezależnie od przeszłości! Ponieważ z jest ustalone, będziemy odtąd pomijali górny indeks przy $T = T^z$. Zdefiniujmy kolejne momenty odnowienia, czyli czasy odwiedzin stanu z:

$$\begin{aligned} T = T_1 &= \min\{n > 0 : X_n = z\}, \\ T_k &= \min\{n > T_{k-1} : X_n = z\}. \end{aligned}$$

Momenty $0 < T_1 < \cdots < T_k < \cdots$ dzielą trajektorię łańcucha na następujące „losowe wycieczki", czyli losowej długości ciągi zmiennych losowych:

$$\begin{array}{ccccc} \underbrace{X_0, \ldots, X_{T_1-1}}_{T_1}, & \underbrace{X_{T_1}, \ldots, X_{T_2-1}}_{T_2-T_1}, & \underbrace{X_{T_2}, \ldots, X_{T_3-1}}_{T_3-T_2}, & \cdots \\ & \uparrow & \uparrow & \\ & X_{T_1} = z & X_{T_2} = z & \cdots \end{array}$$

Wycieczka zaczyna się w punkcie z i kończy tuż przed powrotem do z. Oznaczmy k-tą wycieczkę symbolem Ξ_k:

$$\begin{aligned} \Xi = \Xi_1 &= (X_0, \ldots, X_{T-1}, T), \\ \Xi_k &= (X_{T_{k-1}}, \ldots, X_{T_k-1}, T_k - T_{k-1}) \end{aligned}$$

Z tego, co powiedzieliśmy wcześniej wynika, że wszystkie „wycieczki" są niezależne. Co więcej wycieczki Ξ_k mają ten sam rozkład, z wyjątkiem być może początkowej, czyli Ξ_1. Jeśli rozkład początkowy jest skupiony w punkcie z, to również wycieczka Ξ_1 ma ten sam rozkład (0 jest wtedy momentem odnowienia).

Podejście regeneracyjne, czyli rozbicie łańcucha na niezależne wycieczki prowadzi do ładnych i łatwych dowodów PWL i CTG dla łańcuchów Markowa. Sformułujemy najpierw pewną wersję *Mocnego Prawa Wielkich Liczb*. Rozważmy funkcję f o wartościach rzeczywistych, określoną na przestrzeni stanów. Przypomnijmy, że $\mathbb{E}_\pi f = \sum_{x\in\mathcal{X}} \pi(x) f(x)$.

Twierdzenie 14.3 (Mocne Twierdzenie Ergodyczne)**.** *Jeśli X_n jest nieprzywiedlnym łańcuchem Markowa, to dla dowolnego rozkładu początkowego ξ i każdej funkcji $f : \mathcal{X} \to \mathbb{R}$,*

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_n) \longrightarrow \mathbb{E}_\pi f \qquad (n \to \infty)$$

z prawdopodobieństwem 1.

Dowód. Zdefiniujmy sumy blokowe:

$$\begin{aligned} \Xi_0(f) &= \sum_{i=0}^{T-1} f(X_i), \\ \Xi_k(f) &= \sum_{i=T_k}^{T_{k+1}-1} f(X_i). \end{aligned}$$

Niech $N(n) = \max\{k : T_k \leqslant n\}$, czyli $T_{N(n)}$ jest ostatnią regeneracją przed momentem n:

$$\begin{array}{ccccccc} 0, \ldots, T_1 - 1, & T_1, \ldots\ldots, & T_{N(n)}, \ldots, n, & \ldots, T_{N(n)+1} - 1, & T_{N(n)+1}, \ldots \\ & \uparrow & \uparrow \quad \uparrow & & \uparrow & \\ & X = z & X = z \quad \bullet & & X = z & \end{array}$$

Oczywiście,

$$T_{N(n)} \leqslant n < T_{N(n)+1}. \tag{14.2}$$

Wiemy, że $\mathbb{E}_z T = m < \infty$. Ponieważ T_k jest sumą k niezależnych zmiennych losowych (długości wycieczek) to wnioskujemy, ze $T_k/k \to m$ z prawdopodobieństwem 1, na mocy *zwykłego Prawa Wielkich Liczb*. Rzecz jasna, tak samo $T_{k+1}/k \to m$. Podzielmy nierówność (14.2) stronami przez $N(n)$ i przejdźmy do granicy (korzystając z tego, że $N(n) \to \infty$ prawie na pewno). Twierdzenie o trzech ciągach pozwala wywnioskować, że

$$\frac{N(n)}{n} \to \frac{1}{m} \text{ p.n.}$$

Załóżmy teraz, że $f \geqslant 0$ i powtórzmy bardzo podobne rozumowanie dla sum

$$\sum_{j=1}^{N(n)} \Xi_j(f) \leqslant S_n(f) = \sum_{i=0}^{n-1} f(X_i) \leqslant \sum_{j=1}^{N(n)+1} \Xi_j(f). \tag{14.3}$$

Po lewej i po prawej stronie mamy sumy niezależnych składników $\Xi_j(f)$. Korzystamy z PWL dla niezależnych zmiennych, dzielimy (14.3) stronami przez $N(n)$ i przejchodzimy do granicy. Otrzymujemy

$$\frac{S_n(f)}{N(n)} \to \mathbb{E}_z \Xi(f) \text{ p.n.}$$

a więc

$$\frac{S_n(f)}{n} \to \frac{\mathbb{E}_z \Xi(f)}{m} = \frac{1}{m} \sum_x \alpha(x) f(x) = \sum_x \pi(x) f(x) \text{ p.n.}$$

Ostatnia równość wynika z Twierdzenia Kaca. Przypomnijmy, że $\alpha(x)$ jest „średnim czasem spędzonym w stanie x" podczas pojedynczej wycieczki.

Jeśli funkcja f nie jest nieujemna, to możemy zastosować rozkład $f = f^+ - f^-$ i wykorzystać już udowodniony wynik. □

Na podobnej idei oparty jest „regeneracyjny" dowód Centralnego Twierdzenia Granicznego (istnieją też zupełnie inne dowody).

Twierdzenie 14.4 (Centralne Twierdzenie Graniczne)**.** *Jeśli X_n jest łańcuchem nieprzywiedlnym, to dla dowolnego rozkładu początkowego ξ i każdej funkcji $f : \mathcal{X} \to \mathbb{R}$,*

$$\frac{1}{\sqrt{n}} \left(\sum_{i=0}^{n-1} [f(X_i) - \mathbb{E}_\pi f] \right) \longrightarrow_d \mathrm{N}(0, \sigma^2_{\mathrm{as}}(f)), \quad (n \to \infty).$$

Ponadto, dla dowolnego rozkładu początkowego ξ zachodzi wzór (10.4), czyli $(1/n)\mathrm{Var}_\xi \sum_{i=0}^{n-1} f(X_i) \to \sigma^2_{\mathrm{as}}(f)$ przy $n \to \infty$.

Szkic dowodu. Trochę więcej jest tu technicznych zawiłości niż w dowodzie PWL, wobec tego zdecydowałem się pominąć szczegóły. W istocie, przedstawię tylko bardzo pobieżnie główną ideę. Bez straty ogólności załóżmy, że $\pi^\top f = 0$. Tak jak w dowodzie PWL, sumę $S_n(f) = \sum_{i=0}^{n-1} f(X_i)$ przybliżamy sumą niezależnych składników, które odpowiadają całkowitym wycieczkom: $S_n(f) \simeq S_{T_{N(n)}}(f) = \sum_{j=1}^{N(n)} \Xi_j(f)$. Ze zwykłego CTG dla niezależnych zmiennych o jednakowym rozkładzie otrzymujemy

$$\frac{1}{\sqrt{k}} \sum_{j=1}^{k} \Xi_j(f) \to_{\mathrm{d}} \mathrm{N}(0, \mathrm{Var}_z \Xi(f)).$$

Jeśli „podstawimy" w miejsce k zmienną losową $N(n)$ i wykorzystamy fakt, że $N(n) \simeq n/m$ (PWL gwarantuje, że $N(n)/n \to 1/m$), to nie powinien dziwić następujący wniosek:

$$\frac{1}{\sqrt{n}} S_n(f) \to_{\mathrm{d}} \mathrm{N}(0, \mathrm{Var}_z \Xi(f)/m).$$

W ten sposób „otrzymujemy" tezę. □

Przytoczony powyżej rozumowanie daje ciekawe przedstawienie asymptotycznej wariancji:

$$\sigma^2_{\mathrm{as}}(f) = \mathrm{Var}_z \Xi(f)/\mathbb{E}_z T. \tag{14.4}$$

Istnieje wiele wyrażeń na asymptotyczną wariancję, przy tym różne wzory wymagają różnych założeń. Najbardziej znany jest wzór (10.5). Sformułujemy w jawny sposób potrzebne założenia, i przepiszemy ten wzór w postaci macierzowej. Najpierw musimy wprowadzić jeszcze parę nowych oznaczeń. Wygodnie będzie utożsamić funkcję f z wektorem kolumnowym

$$f = \begin{pmatrix} f(1) \\ \vdots \\ f(d) \end{pmatrix}.$$

Niech

$$\Pi = \mathrm{diag}(\pi(1), \ldots, \pi(d)),$$

gdzie π oznacza, jak zwykle, rozkład stacjonarny. Możemy teraz napisać

$$\mathbb{E}_\pi f(X_0) = \sum_x \pi(x) f(x) = \pi^\top f$$

oraz

$$\begin{aligned} \mathrm{Var}_\pi f(X_0) &= \mathbb{E}_\pi f^2(X_0) - [\ \mathbb{E}_\pi f(X_0)\ ]^2 \\ &= f^\top \Pi f - f^T \pi\pi^T f \\ &= f^\top \Pi (I - 1\pi^\top) f. \end{aligned}$$

Podobnie,

$$\begin{aligned} \mathrm{Cov}_\pi[\ f(X_0), f(X_n)\ ] &= \mathbb{E}_\pi[\ f(X_0) f(X_m)\ ] - [\mathbb{E}_\pi f(X_0)\ ]^2 \\ &= f^\top \Pi P^n f - f^\top \pi\pi^\top f \\ &= f^\top \Pi (P^n - 1\pi^\top) f. \end{aligned}$$

W powyższym wzorze, I jest macierzą identycznościową, 1 oznacza kolumnę jedynek. Latwo sprawdzić, że $P^n - 1\pi^\top = (P - 1\pi^\top)^n$ dla $n > 0$ (ale nie dla $n = 0$). Jeśli P jest macierzą przejścia nieprzywiedlnego i nieokresowego łańcucha Markowa, to $P^n - 1\pi^\top \to 0$ przy $n \to \infty$ na mocy Słabego Twierdzenia Ergodycznego. Stąd wynika zbieżność następujących szeregów:

$$\begin{aligned} A &= \sum_{n=0}^{\infty} (P^n - 1\pi^\top), \\ Z &= \sum_{n=0}^{\infty} (P - 1\pi^\top)^n = \left(I - P + 1\pi^\top\right)^{-1}. \end{aligned}$$

Macierz Z nazywamy *macierzą fundamentalną*. Ponieważ $P^0 - 1\pi^\top = I - 1\pi^\top$ zaś $(P - 1\pi^\top)^0 = I$ więc $Z = A + 1\pi^\top$.

Stwierdzenie 14.1 (Asymptotyczna wariancja)**.** *Jeśli łańcuch jest nieprzywiedlny i nieokresowy, to zachodzi wzór* (10.5)*, czyli*

$$\sigma^2_{\text{as}}(f) = \sigma^2(f) \sum_{n=-\infty}^{\infty} \rho_n(f),$$

gdzie $\sigma^2(f) = \operatorname{Var}_\pi f(X_0)$ *i* $\rho_n(f) = \operatorname{corr}_\pi [f(X_0), f(X_n)]$*, W postaci macierzowej asymptotyczna wariancja wyraża się następująco*

$$\sigma^2_{\text{as}}(f) = f^\top (2\Pi Z - \pi\pi^\top - \Pi) f = f^\top (2\Pi A + \pi\pi^\top - \Pi) f.$$

Szkic dowodu. Załóżmy, że rozkładem początkowym jest π i skorzystamy ze stacjonarności łańcucha:

$$\begin{aligned}
\frac{1}{n}\operatorname{Var}_\pi \left(\sum_{i=0}^{n-1} f(X_i) \right) &= \frac{1}{n} \sum_{i=0}^{n-1} \operatorname{Var}_\pi(X_i) + \frac{2}{n} \sum_{i=0}^{n-1} \sum_{j=i}^{n-1} \operatorname{Cov}_\pi(f(X_i), f(X_j)) \\
&= \operatorname{Var}_\pi f(X_0) + 2 \sum_{k=1}^{n-1} \frac{n-k}{n} \operatorname{Cov}_\pi(f(X_0), f(X_k)) \\
&\to \operatorname{Var}_\pi f(X_0) + 2 \sum_{k=1}^{\infty} \operatorname{Cov}_\pi(f(X_0), f(X_k)), \quad (n \to \infty).
\end{aligned}$$

Poprawność przejścia do granicy w ostatniej linijce wynika z elementarnego faktu, że dla dowolnego ciągu liczbowego a_n mamy $\lim_{n\to\infty} \sum_{k=1}^{n-1} \frac{n-k}{n} a_k = \sum_{k=1}^{\infty} a_k$, o ile szereg po prawej stronie równości jest zbieżny. To zaś, przy założeniu nieprzywiedlności i nieokresowości, wynika ze STE (skorzystaliśmy już z tego faktu uzasadniając poprawność definicji Z i A).

Pokazaliśmy w ten sposób, że $(1/n)\operatorname{Var}_\pi \sum_{i=0}^{n-1} f(X_i)$ zmierza do granicy, która jest równa prawej stronie wzoru (10.5). Pominiemy uzasadnienie, że można zastąpić rozkład stacjonarny π przez dowolny rozkład początkowy ξ oraz że wzór (10.5) definiuje tę samą wielkość co (14.4).

Macierzowe wyrażenia $\sigma^2_{\text{as}}(f)$ wynikają ze wzorów na kowariancje oraz z określenia macierzy A i Z. □

Zauważmy, że ani CTG ani PWL nie wymagały założenia o nieokresowości ale w 14.1 to założenie jest potrzebne.

Jak widać, asymptotyczna wariancja wyraża się w postaci formy kwadratowej $\sigma^2_{\text{as}}(f) = f^\top C f$ o współczynnikach

$$\begin{aligned}
C(x,y) &= \pi(x)A(x,y) + A(y,x)\pi(y) + \pi(x)\pi(y) - \pi(x)\mathbb{I}(x=y), \\
&= \pi(x)Z(x,y) + Z(y,x)\pi(y) - \pi(x)\pi(y) - \pi(x)\mathbb{I}(x=y).
\end{aligned}$$

Okazuje się, że „asymptotyczne obiążenie" estymatora $\hat\theta_n = (1/n)\sum_{i=0}^{n-1} f(X_i)$ wartości oczekiwanej $\theta = \mathbb{E}_\pi f$ można napisać w postaci formy dwuliniowej $\xi^\top A f$, gdzie ξ jest rozkładem początkowym. Podsumowując,

$$\begin{aligned}
\operatorname{Var}_\xi(\hat\theta_n) &= \frac{1}{n} f^\top C f + o\left(\frac{1}{n}\right), \\
\mathbb{E}_\xi \hat\theta_n - \theta f &= \frac{1}{n} \xi^\top A f + o\left(\frac{1}{n}\right).
\end{aligned}$$

Uzasadnienie drugiej części powyższego wzoru pozostawiam jako ćwiczenie. Stąd z łatwością otrzymujemy ważny wzór (10.6) z Rozdziału 10 (wyrażenie na błąd średniokwadratowy).

14.3. Lańcuchy sprzężone i zbieżność rozkładów

Dowód Słabego Twierdzenia Ergodycznego, który przedstawimy, opiera się na tak zwanym sprzęganiu (ang. *coupling*), czyli metodzie „dwóch cząstek". Ta metoda, jak się okaże, nie tylko pozwala udowodnić zbieżność rozkładów prawdopodobieństwa, ale daje w wielu przypadkach bardzo dobre oszacowania *szybkości zbieżności*.

14.3.1. Odległość pełnego wahania

Najpierw zajmiemy się określeniem odległości między rozkładami. Dla naszych celów najbardziej przydatna będzie następująca metryka. Niech ν i λ będą dwoma rozkładami prawdopodobieństwa na skończonej przestrzeni $\mathcal{X}$. Odległość *pełnego wahania* pomiędzy ν i λ określamy wzorem

$$\|\nu - \lambda\|_{\mathrm{tv}} = \max_{\mathcal{A}\subseteq\mathcal{X}} |\nu(\mathcal{A}) - \lambda(\mathcal{A})|. \tag{14.5}$$

Jak zwykle, możemy utożsamić rozkład prawdopodobieństwa na $\mathcal{X}$ z funkcją, przypisującą prawdopodobieństwa pojedynczym punktom $x \in \mathcal{X}$. Zauważmy, że

$$\|\nu - \lambda\|_{\mathrm{tv}} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\nu(x) - \lambda(x)|. \tag{14.6}$$

Istotnie, ponieważ rozpatrujemy dwie miary probabilistyczne, dla których $\nu(\mathcal{X}) = \lambda(\mathcal{X}) = 1$, więc $\|\nu-\lambda\| = \nu(\mathcal{B})-\lambda(\mathcal{B})$ dla $\mathcal{B} = \{x : \nu(x) > \lambda(x)\}$. Ale $\sum_{x\in\mathcal{B}}(\nu(x)-\lambda(x)) = \sum_{x\in\mathcal{X}\setminus\mathcal{B}}(\lambda(x)-\nu(x)) = \frac{1}{2}\sum_{x\in\mathcal{X}}|\nu(x)-\lambda(x)|$.

Dla zmiennej losowej $X : \Omega \to \mathcal{X}$ napis $X \sim \nu$ będzie oznaczał fakt, że X ma rozkład prawdopodobieństwa ν, czyli $\mathbb{P}(X = x) = \nu(x)$,

Lemat 14.2. *Jeżeli $X, Y : \Omega \to \mathcal{X}$ są dwiema zmiennymi losowymi określonymi na tej samej przestrzeni probabilistycznej i $X \sim \nu$ i $Y \sim \lambda$, to*

$$\|\nu - \lambda\|_{\mathrm{tv}} \leqslant \mathbb{P}(X \neq Y).$$

Dowód. Niech $d = \mathbb{P}(X \neq Y)$. Dla dowolnego $\mathcal{A} \subseteq \mathcal{X}$ mamy

$$\nu(\mathcal{A}) = \mathbb{P}(X \in \mathcal{A}) \leqslant \mathbb{P}(Y \in \mathcal{A}) + \mathbb{P}(X \neq Y) = \lambda(\mathcal{A}) + d.$$

Symetrycznie, $\nu(\mathcal{A}) \leqslant \lambda(\mathcal{A}) + d$. Zatem $\|\nu - \lambda\|_{\mathrm{tv}} \leqslant d$. □

Interesujące jest, że Lemat 14.2 daje się, w pewnym sensie, odwrócić. Co prawda, to nie będzie potrzebne w dowodzie Słabego Twioerdzenia Ergodycznego, ale później okaże się bardzo pomocne.

Lemat 14.3. *Jeżeli ν i λ są rozkładami prawdopodobieństwa na $\mathcal{X}$, to istnieją zmienne losowe X i Y określone na tej samej przestrzeni probabilistycznej, takie, że $X \sim \nu$ i $Y \sim \lambda$ i*

$$\|\nu - \lambda\|_{\mathrm{tv}} = \mathbb{P}(X \neq Y).$$

Dowód. Niech $\|\nu - \lambda\|_{\mathrm{tv}} = d$. Bez straty ogólności możemy przyjąć, że X i Y są zmiennymi losowymi określonymi na przestrzeni probabilistycznej $\Omega = \mathcal{X}\times\mathcal{X}$. Należy podać *łączny* rozkład zmiennych losowych X i Y, czyli miarę probabilistyczną χ na $\mathcal{X}\times\mathcal{X}$ taką, że $\sum_y \chi(x,y) = \nu(x)$, $\sum_x \chi(x,y) = \lambda(y)$ i $\sum_x \chi(x,x) = 1 - d$.

Niech

$$\chi(x,x) = \min(\nu(x), \lambda(x)) = \begin{cases} \nu(x) & \text{dla } x \in \mathcal{A}; \\ \lambda(x) & \text{dla } x \in \mathcal{B}, \end{cases}$$

gdzie $\mathcal{A} = \{x : \nu(x) \leqslant \lambda(x)\}$ i $\mathcal{B} = \{x : \nu(x) > \lambda(x)\}$.

Mamy oczywiście $d = 1 - \sum_x \chi(x,x)$ i jest jasne, że tabelka łącznego rozkładu $\chi(x,y) = \mathbb{P}(X = x, Y = y)$ musi być postaci macierzy blokowej

$$\begin{matrix} x\in\mathcal{A} \left\{ \right. \\ x\in\mathcal{B} \left\{ \right. \end{matrix} \begin{matrix} D_{\mathcal{A}} & 0 \\ G & D_{\mathcal{B}} \\ \underbrace{}_{y\in\mathcal{A}} & \underbrace{}_{y\in\mathcal{B}} \end{matrix},$$

gdzie $D_{\mathcal{A}}$ i $D_{\mathcal{B}}$ są macierzami diagonalnymi. Pozostaje tylko odpowiednio „rozmieścić pozostałą masę prawdopodobieństwa" d w macierzy G. Możemy na przykład przyjąć, dla $x \in \mathcal{B}$ i $y \in \mathcal{A}$,

$$\chi(x,y) = \frac{1}{d}\left(\nu(x) - \lambda(x)\right)\left(\lambda(y) - \nu(y)\right).$$

Mamy wtedy $\sum_{y\in\mathcal{A}} \chi(x,y) = \nu(x) - \lambda(x)$, więc $\sum_y \chi(x,y) = \nu(x)$ dla $x \in \mathcal{B}$ i podobnie $\sum_x \chi(x,y) = \lambda(x)$ dla $y \in \mathcal{A}$. Określony przez nas rozkład łączny χ ma więc masę $1 - d$ na przekątnej i żądane rozkłady brzegowe. □

14.3.2. Sprzęganie

Rozważmy „podwójny" łańcuch Markowa (X_n, X'_n) na przestrzeni stanów $\mathcal{X} \times \mathcal{X}$. Przypuśćmy, że każda z dwóch „współrzędnych", oddzielnie rozpatrywana, jest łańcuchem o macierzy przejścia P. Mówiąc dokładniej, zakładamy, że

$$\begin{aligned}\mathbb{P}\left(X_{n+1} = y, X'_{n+1} = y' | X_n = x, X'_n = x', X_{n-1}, X'_{n-1}, \ldots, X_0, X'_0\right) \\ = \bar{P}\big((x,x'),(y,y')\big),\end{aligned}$$

gdzie macierz przejścia $\bar{P}$ podwójnego łańcucha spełnia następujące warunki:

$$\begin{aligned}\sum_{y'} \bar{P}\big((x,x'),(y,y')\big) &= P(x,y) \quad \text{dla każdego } x' \\ \sum_{y} \bar{P}\big((x,x'),(y,y')\big) &= P(x',y') \quad \text{dla każdego } x.\end{aligned} \tag{14.7}$$

Widać, że $X_0, X_1, \ldots, X_n, \ldots$ jest łańcuchem Markowa z prawdopodobieństwami przejścia P i to samo można powiedzieć o $X'_0, X'_1, \ldots, X'_n, \ldots$. Załóżmy ponadto, że od momentu, gdy oba łańcuchy się spotkają, dalej „poruszają się" już razem. Innymi słowy,

$$\bar{P}\big((x,x),(y,y')\big) = \begin{cases} P(x,y) & \text{jeśli } y = y', \\ 0 & \text{jeśli } y \neq y'. \end{cases}$$

Nazwiemy konstrukcję takiej pary *sprzęganiem* łańcuchów (bardziej znany jest angielski termin *coupling*).

Oznaczmy przez T moment spotkania się łańcuchów:

$$T = \min\{n > 0 : X_n = X'_n\}. \tag{14.8}$$

Podstawową rolę odgrywa następujące spostrzeżenie:

$$\|\mathbb{P}(X_n \in \cdot) - \mathbb{P}(X_n \in \cdot)\|_{\mathrm{tv}} \leqslant \mathbb{P}(X_n \neq X'_n) = \mathbb{P}(T > n).$$

Jeśli teraz łańcuch X'_n „wystartuje" z rozkładu stacjonarnego, czyli $X_0 \sim \pi$ to $X_n \sim \pi$ dla każdego n i otrzymujemy

$$\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\|_{\mathrm{tv}} \leqslant \mathbb{P}(T > n). \tag{14.9}$$

Aby udowodnić zbieżność $\mathbb{P}(X_n \in \cdot) \to \pi(\cdot)$ wystarczy skonstruować parę łańcuchów sprzężonych, które się spotkają z prawdopodobieństwem 1: $\mathbb{P}(T < \infty) = 1$. Możemy teraz udowodnić (10.1), przynajmniej dla łańcuchów na skończonej przestrzeni stanów.

Twierdzenie 14.5 (Słabe Twierdzenie Ergodyczne)**.** *Jeśli łańcuch Markowa na skończonej przestrzeni stanów jest nieprzywiedlny i nieokresowy, to*

$$\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\|_{\mathrm{tv}} \to 0.$$

Dowód. Rozważmy parę łańcuchów sprzężonych, które poruszają się *niezależnie* aż do momentu spotkania.

$$\bar{P}((x,x'),(y,y')) = \begin{cases} P(x,y)P(x',y') & \text{jeśli } x \neq x', \\ P(x,y) & \text{jeśli } x = x' \text{ i } y = y', \\ 0 & \text{jeśli } x = x' \text{ i } y \neq y'. \end{cases} \tag{14.10}$$

Żeby pokazać, że $\mathbb{P}(T < \infty) = 1$ wystarczy zauważyć, że do przed momentem spotkania, łańcuch podwójny ewoluuje zgodnie z prawdopodobieństwami przejścia

$$\tilde{P}((x,x'),(y,y')) = P(x,y)P(x',y').$$

Łańcuch odpowiadający $\tilde{P}$ jest nieprzywiedlny. Istotnie, możemy znaleźć takie n_0, że dla $n \geqslant n_0$ wszystkie elementy macierzy P^n są niezerowe. Stąd $P^n((x,x'),(y,y')) = P^n(x,y)P^n(x',y') > 0$ dla dowolnych x, x', y, y'. Wystarczy teraz powołać się na Wniosek 14.2: podwójny łańcuch z prawdopodobieństwem 1 prędzej czy później dojdzie do każdego punktu przestrzeni $\mathcal{X} \times \mathcal{X}$, a zatem musi dojść do „przekątnej" $\{(x,x) : x \in \mathcal{X}\}$. □

Uwaga 14.2. W dowodzie Twierdzenia 14.5 wykorzystaliśmy w istotny sposób nieokresowość macierzy przejścia P (dla pojedynczego łańcucha), choć to mogło nie być wyraźnie widoczne. Jeśli P jest nieprzywiedlna ale okresowa, wtedy $\tilde{P}$ jest nieprzywiedlna. Na przykład, niech $\mathcal{X} = \{0,1\}$ i

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Wtedy, oczywiście, $\tilde{P}^n((0,0),(0,1)) = 0$ bo $P^n(0,0) = 0$ dla nieparzystych n zaś $P^n(0,1) = 0$ dla parzystych n.

Ten sam trywialny przykład pokazuje, że dla łańcuchów okresowych teza Słabego Twierdzenia Ergodycznego nie jest prawdziwa.

W istocie, przytoczony przez nas dowód Twierdzenia 14.5 daje nieco więcej, niż tylko zbieżność rozkładów. Z Wniosku 14.2 wynika, że

$$\|\mathbb{P}(X_n \in \cdot) - \pi(\cdot)\|_{\mathrm{tv}} \leqslant c\gamma^n$$

dla pewnych stałych $c < \infty$ i $\gamma < 1$. Dla naszych celów takie ogólnikowe stwierdzenie nie jest wystarczające. Skupimy się na przykładach łańcuchów używanych w algorytmach MCMC,

dla których znajdziemy jawne oszacowania, z konkretnymi stałymi. Zobaczymy, że użycie niezależnych kopii łańcucha w dowodzie Twierdzenia 14.5, wzór (14.10) jest konstrukcją dalece nieoptymalną. W wielu przykładach istnieją łańcuchy sprzężone znacznie szybciej „zmierzające do spotkania".

Przykład 14.1 (Błądzenie po kostce)**.** Niech $\mathcal{X} = \{0,1\}^n$ i $\pi = \mathrm{U}(\mathcal{X})$, czyli $\pi(x) = 1/2^n$ dla każdego x. Rozważmy łańcuch Markowa X_n, którego krok polega na wylosowaniu jednej, losowo wybranej współrzędnej z rozkładu $(1/2, 1/2)$ na zbiorze $\{0,1\}$ i pozostawieniu pozostałych współrzędnych bez zmian. Formalnie,

$$P(x,y) = \frac{1}{2d}\sum_{i=1}^{d} \mathbb{I}(x_{-i} = y_{-i}).$$

Jest to zatem „losowe błądzenie" wzdłuż krawędzi n-wymiarowej kostki lub inaczej próbnik Gibbsa. Rzecz jasna, dokładne genrowanie z rozkładu jednostajnego na kostce jest łatwe i nie ptrzebujemy do tego łańcuchów Markowa, ale nie o to teraz chodzi. Chcemy zilustrować jak metoda sprzęgania pozwala oszacować szybkość zbieżności łańcucha na możliwie prostym przykładzie. Skonstruujmy parę łańcuchów sprzężonych w taki sposób: wybieramy współrzędną i oraz losujemy jej nową wartość z rozkładu $(1/2, 1/2)$ po czym zmieniamy w ten sam sposób obie kopie. Formalnie,

$$\bar{P}((x,x'),(y,y')) = \frac{1}{2d}\sum_{i=1}^{d} \mathbb{I}(x_{-i} = y_{-i}, x'_{-i} = y'_{-i}, y_i = y'_i).$$

Jest jasne, że to jest poprawny *coupling* (sprzęganie), to znaczy spełnione są równania (14.10). Spotkanie obu kopii nastąpi *najpóźniej* w momencie gdy każda ze współrzędnych zostanie wybrana przynajmniej raz. Zatem

$$\mathbb{P}(T > n) \leqslant \left(1 - \frac{1}{d}\right)^n.$$

Nie trudno wyobrazić sobie, że dla *niezależnego* couplingu określonego wzorem (14.10), czasu oczekiwania na spotkanie obu kopii jest na ogół dużo, dużo dłuższy.

15. Markowowskie Monte Carlo VI. Oszacowania dokładności

W tym rozdziale zajmiemy się problemem oszacowania błędu markowowskich algorytmów Monte Carlo. W odróżnieniu od wstępnych rozważań w Rozdziale 10, będą nas interesowały ścisłe nierówności, a nie oceny oparte na twierdzeniach granicznych. Skupimy się na rozkładzie spektralnym macierzy przejścia. Jest to najbardziej znana i zapewne najskuteczniejsza metoda otrzymywania dobrych oszacowań, przynajmniej dla łańcuchów odwracalnych na przestrzeni skończonej.

15.1. Reprezentacja spektralna macierzy odwracalnej

Rozważmy łańcuch nieprzywiedlny i odwracalny z macierzą przejścia P i rozkładem stacjonarnym $\pi^\top$. Zakładamy więc, że dla dowolnych $x, y \in \mathcal{X}$ spełniona jest zależność

$$\pi(x)P(x,y) = \pi(y)P(y,x).$$

Niech

$$\Pi = \operatorname{diag}(\pi) = \begin{pmatrix} \pi(1) & 0 & \cdots & 0 \\ 0 & \pi(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi(d) \end{pmatrix}.$$

Warunek odwracalności możemy zapisać w postaci $\Pi P = P^\top \Pi$. Wyposażmy przestrzeń $\mathbb{R}^d$ w iloczyn skalarny

$$\langle f, g\rangle_\pi = f^\top \Pi g = \sum_x f(x)g(x)\pi(x),$$

gdzie $f, g \in \mathbb{R}$ traktujemy jak wektory kolumnowe. Oczywiście, norma funkcji f jest zdefiniowana wzorem $\|f\|_\pi{}^2 = \langle f, f\rangle_\pi$. Macierz P traktujemy jako operator działający z lewej strony na funkcje, z prawej strony na rozkłady prawdopodobieństwa. Zgodnie z regułami mnożenia macierzy i wektorów, wzory na Pf i $\xi^\top P$ są następujące:

$$Pf(x) = \sum_y P(x,y)f(y), \qquad \xi^\top P(y) = \sum_x \xi(x)P(x,y).$$

Odwracalność P implikuje

$$\langle f, Pg\rangle_\pi = f^\top \Pi P g = f^\top P^\top \Pi g = g^\top \Pi P f = \langle Pf, g\rangle_\pi.$$

Znaczy to, że P jest macierzą operatora samosprzężonego względem iloczynu skalarnego $\langle\cdot,\cdot\rangle_\pi$. W skrócie powiemy, że P jest π-samosprzężona. Wartości własne P są rzeczywiste i zawarte w przedziale $[-1, 1]$. Uporządkujmy je w kolejności malejącej:

$$1 = \lambda_0 > \lambda_1 \geqslant \cdots \lambda_{d-1} \geqslant -1.$$

Wiadomo, że $\lambda_0 = 1$ jest pojedynczą wartością własną. Jeśli łańcuch jest nieokresowy, to $\lambda_{d-1} > -1$. Niech

$$\Lambda = \operatorname{diag}(1, \lambda_1, \ldots, \lambda_{d-1}).$$

Reprezentacja spektralna macierzy P jest następująca:

$$P = V \Lambda V^\top \Pi,$$

gdzie

$$V^\top \Pi V = I$$

lub, równoważnie, $VV^\top \Pi = I$. Zauważmy, że kolumny macierzy

$$V = (1, v_1, \ldots, v_{d-1})$$

są prawostronnymi wektorami własnymi macierzy P tworzącymi bazę π-ortonormalną. Mamy więc $Pv_i = \lambda v_i$ (oczywiście, $v_0 = 1$ jest tu wektorem jedynek) oraz

$$\langle v_i, v_j \rangle_\pi = v_i^\top \Pi v_j = \mathbb{I}(i = j).$$

Lewostronne wektory własne P (zapisane wierszowo) są postaci $v_i^\top \Pi$: mamy $v_i^\top \Pi P = \lambda_i v_i^\top \Pi$. W szczególności, $v_0^\top \Pi = \pi^\top$. Zapiszmy reprezentację spektralną P w bardziej jawnej formie:

$$P = \sum_{i \geqslant 0} \lambda_i v_i v_i^\top \Pi,$$

przy tym

$$\sum_{i \geqslant 0} v_i v_i^\top \Pi = I.$$

Zauważmy jeszcze, że pierwszy (lub raczej - zerowy) składnik jest macierzą stabilną (o jednakowych wierszach):

$$v_0 v_0^\top \Pi = 1\pi^\top.$$

Reprezentacja spektralna prowadzi do zgrabnych wyrażeń na potęgi macierzy. Latwo zauważyć, że $P^n = V \Lambda^n V^\top \Pi$, czyli

$$P^n = \sum_{i \geqslant 0} \lambda_i^n v_i v_i^\top \Pi = 1\pi^\top + \sum_{i \geqslant 1} \lambda_i^n v_i v_i^\top \Pi. \tag{15.1}$$

Wiemy, że wszystkie wartości własne λ_i z wyjątkiem zerowej ($\lambda_0 = 1$) oraz, być może, ostatniej ($\lambda_{d-1} \geqslant -1$) są co do modułu mniejsze niż 1. Jeżeli więc $\lambda_{d-1} > -1$, to wszystkie składniki sumy we wzorze (15.1) z wyjątkiem początkowego zmierzają do zera i w rezultacie

$$P^n \to 1\pi^\top \quad (n \to \infty).$$

Jest to nic innego jak teza Słabego Twierdzenia Ergodycznego (Twierdzenie 14.5), otrzymana zupełnie inną metodą, przy założeniu odwracalności. Można pokazać, że warunek $\lambda_{d-1} > -1$ jest równoważny nieokresowości, i jest konieczny (jeśli $\lambda_{d-1} = -1$ to łańcuch ma okres 2).

Następujący lemat będzie podstawą dalszych rozważań i umożliwi „przerobienie" STE na jawne wyniki. Zdefiniujmy

$$\lambda = \max(\lambda_1, \ |\lambda_{d-1}| \ ).$$

Załóżmy, że łańcuch jest nieokresowy, więc $\lambda < 1$. Pokażemy, że operator P ograniczony do podprzestrzeni $\{f : f \perp 1\} \subset \mathbb{R}^d$ ortogonalnej do funkcji stałych jest zwężający, ze stałą $\lambda < 1$.

Lemat 15.1. *Jeżeli* $\langle f, 1 \rangle_\pi = 0$ *to* $\|Pf\|_\pi \leqslant \lambda \|f\|_\pi$.

Dowód. Wystarczy zauważyć, że

$$\begin{aligned}\langle Pf, Pf\rangle_\pi &= \langle f, P^2 f\rangle_\pi \\ &= f^\top \Pi 1\pi^\top f + f^\top \sum_{i\geqslant 1} \lambda_i^{2n} v_i v_i^\top \Pi f \\ &= \sum_{i\geqslant 1} \lambda_i^{2n} \langle v_i, f\rangle_\pi^2 \\ &\leqslant \lambda^{2n} \sum_{i\geqslant 1} \langle v_i, f\rangle_\pi^2 \\ &= \lambda^{2n}\langle f, f\rangle.\end{aligned}$$

Korzystamy tu z faktu, że P jest samosprzężony, ze wzoru (15.1) i z tego, że $f^\top \Pi 1 = 0$. □

15.1.1. Oszacowanie szybkości zbieżności

Przejdźmy teraz do jawnych oszacowań szybkości zbieżności w STE. Wyniki zawarte z tym podrozdziale pochodzą z pracy Diaconisa i Strooka [5]. Dla rozkładu prawdopodobieństwa $\xi^\top$ definiujemy „odległość" χ^2 od rozkładu stacjonarnego wzorem

$$\chi^2(\pi, \xi) = \sum_x \frac{(\ \xi(x) - \pi(x)\ )^2}{\pi(x)} = (\xi^\top - \pi^\top)\Pi^{-1}(\xi - \pi).$$

Ta „odległość" nie ma własności symetrii, więc nie jest metryką, ale to nie przeszkadza. Istotna jest interpretacja χ^2 jako „odstępstwa od stacjonarności". Wykorzystamy podejście spektralne, w szczególności Lemat 15.1. Niech $\chi = \sqrt{\chi^2}$.

Stwierdzenie 15.1 (Diaconis i Strook)**.** *Jeśli łańcuch odwracalny, nieprzywiedlny i nieokresowy ma rozkład początkowy ξ, to*

$$\chi(\pi, P^n\xi) \leqslant \lambda^n \chi(\pi, \xi).$$

Dowód. Zastosujmy Lemat 15.1 do wektora $\Pi^{-1}(\xi - \pi)$ który, jak łatwo zauważyć, jest prostopadły do 1. Otrzymujemy

$$\chi(\pi, P^n\xi) = \|P^n\Pi^{-1}(\xi - \pi)\|_\pi \leqslant \lambda^n \|\Pi^{-1}(\xi - \pi)\|_\pi = \lambda^n \chi(\pi, \xi).$$

□

Wniosek 15.1. *Dla łańcucha o rozkładzie początkowym skupionym w punkcie x,*

$$\chi(\pi, P^n(x, \cdot)) \leqslant \lambda^n \sqrt{\frac{1 - \pi(x)}{\pi(x)}} \leqslant \frac{\lambda^n}{\sqrt{\pi(x)}}.$$

Istotnie, mamy jeszcze jedno wyrażenie na „odległość" χ^2: dla dowolnego rozkładu ξ,

$$\chi^2(\pi, \xi) = \xi^\top \Pi^{-1} \xi - 1.$$

Wystarczy teraz podstawić $\xi(y) = \mathbb{I}(y = x)$, aby otrzymać $\chi^2 = 1/\pi(x) - 1$.

15.1.2. Oszacowanie normy pełnego wahania

Istnieje prosta nierówność pomiędzy normą pełnego wahania i „odległością" χ^2:

$$\|\ \xi - \pi\ \|_{\mathrm{tv}} = \frac{1}{2}\sum_x\ |\ \xi(x) - \pi(x)\ | \leqslant \frac{1}{2}\ \chi(\pi, \xi).$$

Wynika to z następującego rachunku:

$$\begin{aligned} 4 \parallel \xi - \pi \parallel_{\mathrm{tv}}^2 &= \left(\sum_x \frac{\mid \xi(x) - \pi(x) \mid}{\sqrt{\pi(x)}} \sqrt{\pi(x)} \right)^2 \\ &\leqslant \sum_x \frac{(\ \xi(x) - \pi(x)\ )^2}{\pi(x)} \sum_x \pi(x) \qquad [\text{ Cauchy-Schwarz }] \\ &= \chi^2(\pi, \xi). \end{aligned}$$

Stąd natychmiast otrzymujemy wniosek

$$\parallel P_n(x, \cdot) - \pi \parallel_{\mathrm{tv}} \leqslant \frac{1}{2} \lambda^n \sqrt{\frac{1 - \pi(x)}{\pi(x)}} \leqslant \frac{\lambda^n}{2\sqrt{\pi(x)}}.$$

15.1.3. Oszacowanie obciążenia estymatora

Rozważmy zadanie obliczania wartości oczekiwanej $\theta = \mathbb{E}_\pi f = \pi^\top f$ dla pewnej funkcji f. Niech $\bar{f} = f - 1\pi^\top f$ oznacza „scentrowaną" funkcję f. Natychmiast widać, że $\bar{f} \perp 1$ i możemy zastosować Lemat 15.1. Stąd już tylko mały krok do oszacowania różnicy między wartością oczekiwaną $\mathbb{E}_\xi f(X_n) = \xi^\top P^n f$ i wartością stacjonarną θ.

Stwierdzenie 15.2. *Jeśli łańcuch jest odwracalny, nieprzywiedlny, nieokresowy i ma rozkład początkowy ξ, to*

$$\mid \mathbb{E}_\xi f(X_n) - \theta \mid \leqslant \lambda^n \chi(\pi, \xi) \sigma(f),$$

gdzie $\sigma^2(f) = \|\bar{f}\|_\pi^2$ jest wariancją stacjonarną funkcji f.

Dowód. Mamy

$$\begin{aligned} \mid \mathbb{E}_\xi f(X_n) - \theta \mid &= \mid (\xi^\top - \pi^\top) P^n f \mid = \mid (\xi^\top - \pi^\top) P^n \bar{f} \mid \\ &= \mid \langle \Pi^{-1}(\xi - \pi), P^n \bar{f} \rangle_\pi \mid \\ &\leqslant \parallel \langle \Pi^{-1}(\xi - \pi) \|_\pi \ \|P^n \bar{f}\|_\pi \qquad [\text{ Cauchy-Schwarz }] \\ &\leqslant \chi(\pi, \xi) \lambda^n \|\bar{f}\|_\pi \qquad [\text{ Lemat 15.1 }]. \end{aligned}$$

□

Rozważmy teraz naturalny estymator

$$\hat{\theta}_{t,n} = \frac{1}{n} \sum_{i=t}^{t+n-1} f(X_i).$$

Jest to średnia wzdłuż trajektorii łańcucha, długości n i „opóźniona" o t. Idea jest jasna: ignorujemy początkowy odcinek trajektorii długości t (tak zwany okres *burn-in*) aby dać łańcuchowi czas na zbliżenie od rozkładu stacjonarnego. Później obliczmy średnią. W ten sposób redukujemy obciążenie. Precyzuje to następujący wniosek.

Wniosek 15.2. *Dla dowolnego n mamy*

$$\mid \mathbb{E}_\xi \hat{\theta}_{t,n} - \theta \mid \leqslant \frac{\lambda^t}{1 - \lambda} \chi(\pi, \xi) \sigma(f).$$

Wynika to z nierówności trójkąta i wzoru na sumę szeregu geometrycznego:

$$\mid \mathbb{E}_\xi \hat{\theta}_{t,n} - \theta \mid \leqslant \sum_{i=t}^{t+n-1} \mid \mathbb{E}_\xi f(X_i) - \theta \mid \leqslant \sum_{i=t}^{\infty} \lambda^i \chi(\pi, \xi) \sigma(f).$$

Zwróćmy uwagę, że obciążenie maleje w tempie geometrycznym przy $t \to \infty$ ale zachowuje sią zaledwie jak $O(1/n)$ przy ustalonym t i $n \to \infty$ (ponieważ początkowe wyrazy sumy mają na obciążenie wpływ dominujący).

15.2. Oszacowanie błędu średniokwadratowego estymatora

Oszacowanie błędu średniokwadratowego (BŚK) estymatora MCMC jest znaczne trudniejsze i subtelniejsze, niż obciążenia.

15.2.1. Asymptotyczna wariancja

Zacznijmy od wyprowadzenia kolejnego wzoru na asymptotyczną wariancję. Przypomnijmy, że zgodnie ze Stwierdzeniem 14.1,

$$\sigma^2_{\text{as}}(f) = f^\top C f,$$

gdzie

$$\begin{aligned} C &= \Pi(2Z - I - 1\pi^\top) = 2\Pi Z - \Pi - \pi\pi^\top \\ &= \Pi(2A - I + 1\pi^\top) = 2\Pi A - \Pi + \pi\pi^\top. \end{aligned}$$

Skorzystajmy z reprezentacji spektralnej macierzy P:

$$P - 1\pi^\top = \sum_{i\geqslant 1} \lambda_i v_i v_i^\top \Pi = V \left(\begin{array}{c|ccc} 0 & & \cdots & 0 \\ \hline & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{d-1} \end{array}\right) .V^\top \Pi$$

Z definicji macierzy A i wzoru (15.1), ponieważ $\sum_{n=0}^\infty \lambda_i^n = 1/(1-\lambda_i)$, więc

$$A = \sum_{n=0}^\infty (P^n - 1\pi^\top) = V \left(\begin{array}{c|ccc} 0 & & \cdots & 0 \\ \hline & \ddots & \cdots & \\ \vdots & \vdots & 1/(1-\lambda_i) & \vdots \\ 0 & & \cdots & \ddots \end{array}\right) V^\top \Pi.$$

Wreszcie, ponieważ $2/(1-\lambda_i) - 1 = (1+\lambda_i)/(1-\lambda_i)$, więc

$$\Pi(2A - 1 + 1\pi^\top) = \Pi V \left(\begin{array}{c|ccc} 0 & & \cdots & 0 \\ \hline & \ddots & \cdots & \\ \vdots & \vdots & (1+\lambda_i)/(1-\lambda_i) & \vdots \\ 0 & & \cdots & \ddots \end{array}\right) V^\top \Pi.$$

Zauważmy teraz, że wektor $\alpha = V^\top \Pi f$ zawiera współrzędne wektora f w bazie ON złożonej z prawych wektorów własnych: $\alpha_i = v_i^\top \Pi f = \langle\, v_i,\ f\,\rangle_\pi$. Udowodniliśmy w ten sposób następujący fakt.

Stwierdzenie 15.3. *Dla łańcucha odwracalnego, wzór na asymptotyczną wariancję przybiera postać*

$$\sigma^2_{\text{as}}(f) = \sum_{i\geqslant 1} \alpha_i^2 \, \frac{1+\lambda_i}{1-\lambda_i},$$

gdzie $\alpha_i = v_i^\top \Pi f = \langle\, v_i,\ f\,\rangle_\pi$.

Wynika stąd ważna nierówność.

Wniosek 15.3. *Dla łańcucha odwracalnego mamy następujące oszacowanie asymptotycznej wariancji:*

$$\sigma^2_{\text{as}}(f) \leqslant \frac{1+\lambda_1}{1-\lambda_1}\sigma^2(f)$$

gdzie λ_1 jest największą wartością własną mniejszą od 1, a $\sigma^2(f)$ jest wariancją stacjonarną.

Istotnie,

$$\sigma^2_{\text{as}}(f) \leqslant \sum_{i\geqslant 1} \alpha_i^2 \, \frac{1+\lambda_i}{1-\lambda_i} \leqslant \frac{1+\lambda_1}{1-\lambda_1} \sum_{i\geqslant 1} \alpha_i^2,$$

a łatwo widzieć, że $\sum_{i\geqslant 1} \alpha_i^2 = \|\bar{f}\|^2_\pi = \sigma^2(f)$. Zwróćmy uwagę, że we Wniosku 15.3 występuje λ_1, a nie $\lambda = \max(\lambda_1, |\lambda_{d-1}|)$ (największa *co do modułu* wartość własna mniejsza od 1).

Na zakończenie przytoczę jeszcze kilka sugestywnych wzorów.

Uwaga 15.1. Macierz $L = I - P$ jest nazywana laplasjanem. Zauważmy, że

$$L = \sum_{i\geqslant 1} (1-\lambda_i)\; v_i v_i^\top \Pi.$$

Ponieważ

$$A = \sum_{i\geqslant 1} \frac{1}{1-\lambda_i}\; v_i v_i^\top \Pi,$$

uzasadnia to interpretację macierzy A jako „uogólnionej odwrotności" laplasjanu.

Dorzućmy przy okazji jeszcze jedno wyrażenie na asymptotyczną wariancję:

$$\sigma^2_{\text{as}}(f) = \langle\; f,\; (2A - \mathbb{I} + \mathbf{1}\pi^T) f \;\rangle_\pi = \langle\; f,\; Af \;\rangle_\pi + \langle\; Af,\; f \;\rangle_\pi - \sigma^2(f).$$

Jeśli $\pi^\top f = 0$, to $\sigma^2(f) = \operatorname{Var}_\pi f = \langle\; f, f \;\rangle_\pi$ i ostatni wzór możemy przepisać w postaci

$$\sigma^2_{\text{as}}(f) = \sigma^2(f) \left(2 \frac{\langle\; f,\; Af \;\rangle_\pi}{\langle\; f,\; f \;\rangle_\pi} - 1 \right)$$

15.2.2. Oszacowanie BŚK

Wyniki w tym podrozdziale zostały otrzymane przez Aldousa [1]. Pomysł polega na tym, żeby najpierw otrzymać nierówność dla łańcucha stacjonarnego, a potem postarać się o uogólnienie dla łańcucha o dowolnym rozkładzie początkowym. Przypomnijmy oznaczenia $\theta = \mathbb{E}_\pi f$, $\hat{\theta}_n = \sum_{i=0}^{n-1} f(X_i)$.

Stwierdzenie 15.4 (Aldous, 1987)**.** *Dla dla łańcucha nieprzywiedlnego, odwracalnego i stacjonarnego,* MSE_π *można oszacować w następujący sposób:*

$$\mathbb{E}_\pi(\hat{\theta}_n - \theta)^2 \leqslant \frac{1+\tilde{\lambda}}{1-\tilde{\lambda}}\sigma^2(f)$$

gdzie $\tilde{\lambda} = \max(\lambda_1, 0)$, zaś λ_1 jest największą wartością własną mniejszą od 1.

Dowód. Korzystamy ze stacjonarności i z rozkładu spektralnego macierzy P.

$$\begin{aligned}
\mathbb{E}_\pi(\hat\theta_n-\theta)^2 &= \mathbb{E}_\pi\left(\frac{1}{n}\sum_{i=0}^{n-1}\bar f(X_i)\right)^2\\
&= \frac{1}{n^2}\sum_{i=0}^{n-1}\mathbb{E}_\pi\bar f(X_i)^2+\frac{2}{n^2}\sum_{i=0}^{n-1}\sum_{j=i+1}^{n-1}\mathbb{E}_\pi\bar f(X_i)\bar f(X_j)\\
&= \frac{1}{n}\mathbb{E}_\pi\bar f(X_0)^2+\frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)\mathbb{E}_\pi\bar f(X_0)\bar f(X_k)\qquad [k=j-i]\\
&= \frac{1}{n}\langle\bar f,\bar f\rangle_\pi+\frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)\langle\bar f,P^k\bar f\rangle_\pi\\
&= \frac{1}{n}\langle\bar f,\bar f\rangle_\pi+\frac{2}{n^2}\sum_{k=1}^{n-1}(n-k)\sum_{i\geqslant 1}\tilde\lambda_i^k\langle v_i,f\rangle_\pi{}^2\\
&= \frac{1}{n}\langle\bar f,\bar f\rangle_\pi+\frac{2}{n^2}\sum_{i\geqslant 1}\langle v_i,f\rangle_\pi^2\sum_{k=1}^{n-1}(n-k)\lambda_i^k\\
&\leqslant \frac{1}{n}\sigma^2(f)+\frac{2}{n^2}\sum_{\substack{i\geqslant 1\\ \lambda_i>0}}\langle v_i,f\rangle_\pi^2\sum_{k=1}^{n-1}(n-k)\lambda_i^k\qquad [\text{pomijamy składniki dla } \lambda_i<0]\\
&\leqslant \frac{1}{n}\sigma^2(f)+\frac{2}{n^2}\sigma^2(f)\sum_{k=1}^{n-1}(n-k)\tilde\lambda^k\\
&\leqslant \frac{1}{n}\sigma^2(f)+\frac{2}{n}\sigma^2(f)\sum_{k=1}^{n-1}\tilde\lambda^k\\
&\leqslant \frac{2}{n}\sigma^2(f)\left(1+\frac{2\tilde\lambda}{1-\tilde\lambda}\right)=\frac{1}{n}\sigma^2(f)\frac{1+\tilde\lambda}{1-\tilde\lambda}.
\end{aligned}$$

Zwróćmy uwagę na miejsce, w którym pomijamy składniki odpowiadające ujemnym wartościom własnym. Uzasadnienie jest takie, że dla , $\lambda_i<0$, składniki sumy $\sum_{k=1}^{n-1}(n-k)\lambda_i^k$ są naprzemian ujemne i dodatnie, o malejących wartościach bezwzględnych. Stąd wynika, że $\sum_{k=1}^{n-1}(n-k)\lambda_i^k<0$ i można tę sumę w nierówności opuścić. Jest to ciekawe zjawisko: *ujemne wartości własne pomagają*, zmniejszają błąd! Działają podobnie jak zmienne antytetyczne. □

Niech teraz $E(x)$ oznacza błąd średniokwadratowy dla łańcucha startującego z punktu $x\in\mathcal{X}$,

$$e_n(x)=\mathbb{E}_x(\hat\theta_n-\theta)^2.$$

Rozpatrzmy łańcuch o rozkładzie początkowym ξ. Niech

$$\hat\theta_{t,n}=\frac{1}{n}\sum_{i=t}^{t+n-1}f(X_i).$$

będzie średnią długości n obliczany po odrzuceniu t początkowych zmiennych.

Stwierdzenie 15.5. *Dla dla łańcucha nieprzywiedlnego, odwracalnego startującego z dowolnego rozkładu* ξ, MSE_ξ *można oszacować w następujący sposób:*

$$\mathbb{E}_\xi(\hat\theta_{t,n}-\theta)^2\leqslant\frac{1+\tilde\lambda}{1-\tilde\lambda}\sigma^2(f)+\lambda^t\chi(\pi,\xi)\max_x|\bar f(x)|^2$$

gdzie $\tilde\lambda=\max(\lambda_1,0)$ *i* $\lambda=\max(\lambda_1,|\lambda_{d-1}|)$.

Dowód. Na mocy Stwierdzeń 15.4 i 15.2 mamy

$$\begin{aligned}\mathbb{E}_\xi(\hat{\theta}_{t,n}-\theta)^2 = \mathbb{E}_\xi e_n(X_t) &\leqslant \mathbb{E}_\pi e_n + |\mathbb{E}_\xi e_n(X_t) - \mathbb{E}_\pi e_n| \\ &\leqslant \frac{1}{n}\sigma^2(f)\frac{1+\tilde{\lambda}}{1-\tilde{\lambda}} + \lambda^t \chi(\pi,\xi)\sigma(e_n) \\ &\leqslant \frac{1}{n}\sigma^2(f)\frac{1+\tilde{\lambda}}{1-\tilde{\lambda}} + \lambda^t \chi(\pi,\xi)\max_x |\bar{f}(x)|^2,\end{aligned}$$

bo $e_n(y) \leqslant \max_x |\bar{f}(x)|^2$, a więc $\sigma(e_n) \leqslant \max_x |\bar{f}(x)|^2$. □

Nierówność podane przez Aldousa w cytowanej pracy była nieco inna.

Stwierdzenie 15.6 (Aldous, 1987)**.** *Dla dla łańcucha nieprzywiedlnego, odwracalnego startującego z dowolnego rozkładu ξ mamy następujące oszacowanie błędu MSE_ξ:*

$$\mathbb{E}_\xi(\hat{\theta}_{t,n}-\theta)^2 \leqslant \left(1+\frac{\lambda^t}{\pi_*}\right)\frac{1+\tilde{\lambda}}{1-\tilde{\lambda}}\sigma^2(f),$$

gdzie $\pi^* = \min_x \pi(x)$.

Dowód. Istotnie, załóżmy, że łańcuch startuje z deterministycznie wybranego punktu $x \in \mathcal{X}$, czyli $\xi = P(x,\cdot)$. Jeśli otrzymamy oszacowanie niezależne od x, dowód będzie zakończony.

$$\begin{aligned}\mathbb{E}_x(\hat{\theta}_{t,n}-\theta)^2 &= \sum_y P^t(x,y)e_n(y) \\ &= \sum_y P^t(x,y)e_n(y) = \sum_y \frac{P^t(x,y)}{\pi(y)}\pi(y)e_n(y) \\ &\leqslant \left(1+\frac{\lambda^t}{\pi_*}\right)\sum_y \pi(y)e_n(y) = \left(1+\frac{\lambda^t}{\pi_*}\right)\mathrm{MSE}_\pi.\end{aligned}$$

i zastosujmy Stwierdzenie 15.4. Wykorzystaliśmy tu nierówność $P^t(x,y)/\pi(y) \leqslant 1+\lambda^t/\pi_*$, która wynika z rozkładu spektralnego:

$$\begin{aligned}\frac{P^t(x,y)}{\pi(y)} &= P^t\Pi^{-1}(x,y) = 1_x^\top P^t \Pi^{-1} 1_y \\ &= 1_x^\top \sum_{i=0}^{d-1} \lambda_i^t v_i v_i^\top 1_y = 1 + 1_x^\top \sum_{i=1}^{d-1} \lambda_i^t v_i v_i^\top 1_y \\ &\leqslant 1 + \lambda^t \sum_{i=1}^{d-1} |1_x^\top v_i v_i^\top 1_y| \\ &\leqslant 1 + \lambda^t \sqrt{\sum_{i=1}^{d-1}(1_x^\top v_i)^2}\sqrt{\sum_{i=1}^{d-1}(v_i 1_y^\top)^2} \\ &\leqslant 1 + \frac{\lambda^t}{\sqrt{\pi(x)}\sqrt{\pi(y)}} \leqslant 1 + \frac{\lambda^t}{\pi_*}.\end{aligned}$$

W powyższym wzorze symbol $1_x^\top$ oznacza wektor $(0,\ldots,0,1,0,\ldots,0)$, gdzie jedynka stoi na x-tym miejscu. Skorzystaliśmy z nierówności Schwarza. □

Literatura

[1] D. Aldous: On the Markov Chain Simulation Method for Uniform Combinatorial Distributions and Simulated Annealing, *Probability in the Engineering and Informational Science*, pp. 33–45, 1987.

[2] S. Asmussen and P.W. Glynn: *Stochastic Simulation, Algorithms and Analysis*, Springer, 2007.

[3] S. Asmussen: *Ruin Probabilities*, World Scientific, 2000, 2001.

[4] P. Bremaud: *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, Springer Verlag, 1999.

[5] P. Diaconis, D. Strook (1991): *Geometric bounds for eigenvalues of Markov chains*, Annals of Applied Probability 1 (1), 36–61.

[6] S. Geman and D. Geman (1984): Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE-PAMI*, 6, 721–741.

[7] C.J. Geyer (1992): Practical Markov Chain Monte Carlo. *Statistical Science* 7 (4), 473–511.

[8] C.J. Geyer (1995, 2005): Markov chain Monte Carlo Lecture Notes. Dostępne na `www.stat.umn.edu/geyer`.

[9] C.J. Geyer, E.A. Thompson (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Statististical Society B*, 54, 3, 657–699.

[10] W.K. Hastings (1970): Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* 57, 97–109.

[11] M. Jerrum, A. Sinclair (1996): The Markov chain Monte Carlo method: an approach to approximate counting and integration, In *Approximation Algorithms for NP-hard Problems*, (Dorit Hochbaum, ed.), PWS, 1996.

[12] : M. Jerrum (1998): Mathematical foundations of the Markov chain Monte Carlo method, In *Probabilistic Methods for Algorithmic Discrete Mathematics*, Springer 1998.

[13] G. Jones (2004): On the Markov chain Central Limit Theorem, *Probability Surveys* 1, 299–320.

[14] F.K.C. Kingman: *Procesy Poissona*, PWN 2002.

[15] J.S. Liu: *Monte Carlo Strategies in Scientific Computing*, Springer 2004.

[16] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller (1953): Equation of state calculation by fast computing machines, *Journal of Chemical Physics*, 21 (6), 1087–1092.

[17] E. Nummelin (2002): MC's for MCMC'ists. *International Statistical Review*, 70, 215–240.

[18] B.D. Ripley: *Stochastic Simulation*, Wiley & Sons, 1987.

[19] C.P. Robert, G. Casella: *Monte Carlo Statistical Methods*, Springer 2004.

[20] G.O. Roberts, J.S. Rosenthal (2004): General state space Markov chains and MCMC algorithms. *Probability Surveys* 1, 20–71.

[21] J.S. Rosenthal (1995): Rates of convergence for Gibbs sampling for variance component models, *Annals of Statistics* 23, 740–761.

[22] M. Rybiński: *Krótkie wprowadzenie do R dla programistów, z elementami statystyki opisowej*, WMIM UW 2009.

[23] R. Zieliński, R. Wieczorkowski: *Komputerowe generatory liczb losowych*, WNT, Warszawa, 1997.