Rachunek Prawdopodobieństwa

ZADANIE DOMOWE 2

Termin wysyłania (MS Teams): **26 listopada 2025 godz. 23:59**

Zadanie 1. *Kule i urny*, patrz np. rozdział 5 w [MU17]

Jednym z klasycznych modeli probabilistycznych mającym zastosowania w algorytmice jest model kul i urn (ang. *balls and bins*). W modelu tym m kul kolejno wrzucanych jest niezależnie i jednostajnie losowo do n ponumerowanych urn. Wrzucenie m kul do n urn w taki sposób możemy utożsamiać z losową funkcją ze zbioru $\{1, \ldots, m\}$ w zbiór $\{1, \ldots, n\}$ (formalnie, przestrzenią zdarzeń elementarnych jest wówczas zbiór $\Omega_{n,m} = \{1, \ldots, n\}^{\{1,\ldots,m\}}$).

Celem tego zadania jest eksperymentalne wyznaczenie następujących wielkości:

(a) B_n – moment pierwszej kolizji; $B_n = k$, jeśli k-ta z wrzucanych kul jest pierwszą, która trafiła do niepustej urny (**paradoks urodzinowy**, ang. *birthday paradox*),

(b) U_n – liczba pustych urn po wrzuceniu n kul,

(c) C_n – minimalna liczba rzutów, po której w każdej z urn jest co najmniej jedna kula (pierwszy moment, w którym nie ma już pustych urn; **problem kolekcjonera kuponów**, ang. *coupon collector's problem*),

(d) D_n – minimalna liczba rzutów, po której w każdej z urn są co najmniej dwie kule (*the siblings of the coupon collector / coupon collector's brother*),

(e) $D_n - C_n$ – liczba rzutów od momentu C_n potrzebna do tego, żeby w każdej urnie były co najmniej dwie kule.

Zaimplementuj symulacje obejmujące wykonanie dla każdego $n \in \{1000, 2000, \ldots, 100\,000\}$ po $k = 50$ niezależnych powtórzeń eksperymentu wrzucania kul do urn i zapisanie (np. do pliku) wszystkich powyższych statystyk. Pojedyncze powtórzenie eksperymentu może polegać na wrzucaniu kul aż do pierwszego momentu, w którym w każdej z urn są co najmniej dwie kule i zliczaniu „po drodze” badanych statystyk.

Zadbaj o to, aby generator liczb pseudolosowych użyty w symulacjach był „dobry” (tj. miał dobre własności statystyczne). Przykładowo, standardowa implementacja funkcji `rand()` w języku C nie jest dobrym generatorem. Możesz wykorzystać np. generator Mersenne Twister.

Po zakończeniu symulacji, korzystając z zebranych danych, dla każdej z badanych zmiennych losowych (B_n, U_n, C_n, D_n oraz $D_n - C_n$) przedstaw na wykresach za pomocą wybranego narzędzia (np. *numpy, Matlab, Mathematica, ...*) wyniki poszczególnych powtórzeń (k punktów danych dla każdego n) oraz ich wartość średnią jako funkcję n. Wyniki poszczególnych prób oraz ich średnie nanieś na wspólny wykres tak, aby można było łatwo określić ich koncentrację wokół wartości średniej.

Dodatkowo wykonaj następujące wykresy (poniżej $b(n)$, $u(n)$, $c(n)$ i $d(n)$ oznaczają, odpowiednio, średnią wartość B_n, U_n, C_n i D_n dla danego n):

(a) iloraz $\frac{b(n)}{n}$ oraz $\frac{b(n)}{\sqrt{n}}$ jako funkcja n,

(b) iloraz $\frac{u(n)}{n}$ jako funkcja n,

(c) iloraz $\frac{c(n)}{n}$, $\frac{c(n)}{n \ln n}$ oraz $\frac{c(n)}{n^2}$ jako funkcja n,

(d) iloraz $\frac{d(n)}{n}$, $\frac{d(n)}{n \ln n}$ oraz $\frac{d(n)}{n^2}$ jako funkcja n,

(e) iloraz $\frac{d(n) - c(n)}{n}$, $\frac{d(n) - c(n)}{n \ln n}$ oraz $\frac{d(n) - c(n)}{n \ln \ln n}$ jako funkcja n.

Uwaga. W ZADANIU DOMOWYM 3. będziemy dalej pracować z modelem kul i urn w kontekście zastosowań algorytmicznych (problem *balanced allocation*). Warto więc przygotować implementację tak, żeby można było ją ponownie wykorzystać (po odpowiednich modyfikacjach). Tam będziemy badać m.in. maksymalną liczbę kul w urnie po wrzuceniu n kul (*maximum load*).

Zadanie 2. Przedstaw wyniki eksperymentów z zadania 1. i odpowiedz na poniższe pytania.

(a) Zaprezentuj wykresy, zwięźle omów uzyskane rezultaty i przedstaw wnioski.

(b) Na podstawie wykresów krótko scharakteryzuj koncentrację wyników uzyskanych w poszczególnych powtórzeniach wokół wartości średniej dla badanych wielkości.

(c) Na podstawie wykresów postaw hipotezy odnośnie asymptotyki wartości średnich badanych wielkości i uzasadnij ich wybór (w razie potrzeby w niektórych wykresach możesz zastosować skalę logarytmiczną).

(d) Zaproponuj rozsądne uzasadnienie użycia nazw *birthday paradox* oraz *coupon collector's problem* pojawiających się w zadaniu 1. (przedstaw stojące za nimi intuicje).

(e) Jakie znaczenie ma *birthday paradox* w kontekście funkcji hashujcych i kryptograficznych funkcji hashujcych?

Rozwiązanie zadania obejmujące

- implementację symulacji (kod źródłowy w wybranym języku programowania) oraz
- uzyskane wyniki i wyciągnięte na ich podstawie wnioski (pdf z wykresami i samodzielnie napisanymi **zwięzłymi** opisami wyników, wnioskami oraz odpowiedziami do zadania 2)

należy przesłać na platformę MS Teams. Nie należy dołączać żadnych zbędnych plików.

Literatura

[MU17] Michael Mitzenmacher and Eli Upfal. *Probability and Computing. Randomization and Probabilistic Techniques in Algorithms and Data Analysis.* Cambridge University Press, 2nd edition, 2017.