Metody Probabilistyczne i Statystyka

ZADANIE DOMOWE 2

Termin wysyłania (MS Teams): **04 grudnia 2022 godz. 23:59**

Za rozwiązanie Zadania 1. i Zadania 2. można uzyskać łącznie **10 pkt**.

Zadanie 1. (*Kule i urny*, patrz np. rozdział 5 w [MU17]) Jednym z klasycznych modeli probabilistycznych, często rozważanym w kontekście zagadnień algorytmicznych, jest model kul i urn (ang. *balls and bins*). W modelu tym m kul wrzucanych jest kolejno do $n \geqslant 1$ ponumerowanych urn. Każda kula wrzucana jest niezależnie z jednakowym prawdopodobieństwem równym $\frac{1}{n}$ do jednej z 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) L_n – maksymalna liczba kul w urnie po wrzuceniu n kul (*maximum load*),

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

(e) 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*),

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

Zaimplementuj symulacje polegające na wykonaniu dla każdego $n \in \{1000, 2000, \ldots, 100\,000\}$ po $k = 50$ niezależnych powtórzeń eksperymentu wrzucania kul do urn i zapisaniu (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" wszystkich 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 np. wykorzystać generator Mersenne Twister.

Po zakończeniu symulacji, korzystając z zebranych danych, dla każdej z badanych statystyk (B_n, U_n, L_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 średnią wartość statystyki jako funkcję n. Wartość średnią oraz wszystkie wyniki poszczególnych prób nanieś na wspólny wykres tak, aby można było łatwo określić ich koncentrację wokół wartości średniej.

Dodatkowo wykonaj następujące wykresy (poniżej $b(n)$, $u(n)$, $l(n)$, $c(n)$ i $d(n)$ oznaczają, odpowiednio, średnią wartość statystyki B_n, U_n, L_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 $\frac{l(n)}{\ln n}$, $\frac{l(n)}{(\ln n)/\ln\ln n}$ oraz $\frac{l(n)}{\ln\ln n}$ jako funkcja n,

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

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

(f) 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.

Zadanie 2. Przedstaw wyniki eksperymentów przeprowadzonych w Zadaniu 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 dla poszczególnych powtórzeń wokół wartości średniej wyznaczanych statystyk.

(c) Przedstaw hipotezy odnośnie asymptotyki wartości średnich badanych statystyk postawione na podstawie analizy wykresów 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
- pdf z wykresami, zwięzłym opisem wyników, wnioskami i 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*. 2nd edition, 2017.