

Metody Probabilistyczne i Statystyka

ZADANIE DOMOWE 2

Termin wysyłania (MS Teams): **26 listopada 2023 godz. 23:59**

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 $n \geq 1$ ponumerowanych urn. Każda kula wrzucana jest niezależnie z jednakowym prawdopodobieństwem równym $1/n$ do jednej z urn. Wrzucenie m kul do n urn w taki sposób możemy utożsamiać z losową funkcją ze zbioru $\{1, \dots, m\}$ w zbiór $\{1, \dots, n\}$ (formalnie, przestrzenią zdarzeń elementarnych jest wówczas zbiór $\Omega_{n,m} = \{1, \dots, n\}^{\{1, \dots, m\}}$).

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 potrzeba do tego, żeby w każdej urnie były co najmniej dwie kule.

Zaimplementuj symulacje polegające na wykonaniu dla każdego $n \in \{1000, 2000, \dots, 100\,000\}$ 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” 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 zmiennych losowych (wielkości 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 zajmować się modelem kul i urn w kontekście pewnych zagadnień algorytmicznych (problem *balanced allocation*). Dlatego warto 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 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 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 hashujących i kryptograficznych funkcji hashujących?

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*. Cambridge University Press, USA, 2nd edition, 2017.