Egzamin z Metod Probabilistycznych w Algorytmice, 02.02.2011

Część II: zadania

Zadanie 1: Haszowanie tabulacyjne [10 punktów] Niech $U = \{0, \ldots, 2^{cr} - 1\}$, dla pewnych $c, r \in \mathbb{N}$, $c > 1$. Innymi słowy, uniwersum U składa się z liczb (cr)-bitowych. Każdą taką liczbę x będziemy reprezentować jako krotkę $(x_1, x_2, \ldots, x_c)$ liczb r-bitowych. (W typowych zastosowaniach U składa się z 32- lub 64-bitowych liczb naturalnych, natomiast $r = 8$, czyli krotka to kolejne bajty liczby x).

Niech $m = 2^k$. Oznaczmy $[\ell] = \{0, \ldots, \ell - 1\}$. Rozważmy następującą rodzinę funkcji haszujących $h : U \to [m]$:

$$H = \{x \mapsto T_1(x_1) \oplus T_2(x_2) \oplus \ldots \oplus T_c(x_c) \mid T_1, ..., T_c : [2^r] \to [m]\},$$

gdzie $\oplus$ oznacza operację xor, np $6 \oplus 3 = (110)_2 \oplus (011)_2 = (101)_2 = 5$.

Innymi słowy, losową funkcję haszującą h z tej rodziny reprezentuje c losowych funkcji, których argumentami są liczby r-bitowe, a wartościami liczby z zakresu $[m]$ (każdą z nich reprezentujemy np. jako tablicę długości 2^r).

Pokaż, że rodzina H jest 3-niezależna, ale nie jest 4-niezależna.

Uwaga: pewną liczbę punktów mozna zdobyć pokazując, że H jest a-niezależna ale nie b-niezależna, dla pewnych (konkretnych) liczb $a \leq 3$, $b \geq 4$ (im bliższych wartościom odpowiednio 3, 4 tym lepiej).

Uwaga natury poznawczej. W 2010r Patrascu i Thorup udowodnili, że rodzina H pomimo niskiego stopnia niezależności dobrze się sprawdza w zastosowaniach takich jak m.in. haszowanie łańcuchowe, haszowanie liniowe, statyczne haszowanie kukułkowe. W szczególności, oczekiwane czasy operacji słownikowych przy haszowaniu liniowym są stałe. Jest to zapewne najbardziej praktyczna z rodzin funkcji haszujących, których jakość została potwierdzona formalnymi dowodami.

Szkic rozwiązania Najpierw pokażemy, że nie ma 4-niezależności. Weźmy $c = 2$ oraz 4 klucze: $x = (0, 0), y = (0, 1), z = (1, 0), u = (1, 1)$. Obliczmy $p = P[h(x) = a \wedge h(y) = b \wedge h(z) = c \wedge h(u) = d]$. Nasze zdarzenie jest równoważne układowi równań:

$$\begin{cases} T_1[0] \oplus T_2[0] = a \\ T_1[0] \oplus T_2[1] = b \\ T_1[1] \oplus T_2[0] = c \\ T_1[1] \oplus T_2[1] = d \end{cases}$$

A więc nasze zdarzenie pociąga za sobą, że $a \oplus b \oplus c \oplus d = 0$, czyli jeśli ten warunek nie zachodzi to $p = 0$ (a powinno być $p = \frac{1}{m^4}$ niezależnie od a, b, c, d).

Teraz 3-niezależność. Rozważmy 3 parami różne klucze x, y, z. r-bitowe bloki kluczy będziemy nazywać bajtami. Musi istnieć taki bajt, który nie jest identyczny dla wszystkich trzech kluczy. A więc dla jednego z kluczy (przez symetrię przyjmijmy, że dla x, i że jest to pierwszy bajt) ten bajt jest inny niż dla pozostałych, tzn $x_1 \neq y_1$ i $x_1 \neq z_1$. To powoduje, że $h(x)$ jest łącznie niezależne od $h(y)$ i $h(z)$, bo po wylosowaniu wartości w tablicach T_j potrzebnych aby ustalić $h(y)$ i $h(z)$ możemy z takim samym prawdopodobieństwem dostać każdą wartość $h(x)$ losując wartość $T_1[x_1]$. Dostaliśmy więc $P[h(x) = a \wedge h(y) = b \wedge h(z) = c] = P[h(x) = a \mid h(y) = b \wedge h(z) = c] \cdot P[h(y) = b \wedge h(z) = c] = \frac{1}{m} \cdot P[h(y) = b \wedge h(z) = c]$. Wystarczy pokazać, że $h(y)$ i $h(z)$ są niezależne. Ale to wynika z tego, że mają bajt, na którym się różnią (powiedzmy bajt nr 1: wtedy po wylosowaniu wartości w tablicach T_j dla pozostałych bajtów możemy z takim samym prawdopodobieństwem dostać dowolne wartości $h(y)$ i $h(z)$ losując $T_1[y_1]$ i $T_1[z_1]$). Czyli $P[h(y) = b \wedge h(z) = c] = \frac{1}{m^2}$, i z wcześniejszych rozważań $P[h(x) = a \wedge h(y) = b \wedge h(z) = c] = \frac{1}{m^3}$ i koniec.

Zadanie 2: Chaos w eterze [10 punktów] n żołnierzy próbuje się jednocześnie skontaktować drogą radiową z jednym z n centrów dowodzenia, z których każde odbiera na unikalnej częstotliwości. Żołnierze mają zsynchronizowane zegarki. Na początku każdej minuty, każdy z żołnierzy, którzy jeszcze nie nawiązali kontaktu wybiera losowo centrum, z którym próbuje się skontaktować. Jeśli dwóch żołnierzy wybierze to samo centrum, to będą się nawzajem zagłuszać i żadnemu z nich nie uda się nawiązać kontaktu. Jeśli natomiast jakiś żołnierz jest jedynym, który kontaktuje się z pewnym centrum, to otrzymuje on dalsze rozkazy i wraca do zabijania. Pozostali żołnierze odczekują minutę, po czym ponawiają próbę z nowo wylosowanym centrum. Udowodnij, że oczekiwana liczba minut potrzebnych, aby wszyscy żołnierze otrzymali rozkazy wynosi $O(\log\log n)$.
Wskazówka Pokaż, że jeśli na początku i-tej minuty kontakt próbowało nawiązać x_i żołnierzy, to wartość oczekiwana liczby żołnierzy, którym się to nie udało nie przekracza $(x_i)^2/n$. Na tej podstawie oszacuj oczekiwaną liczbę minut potrzebną, żeby zostało $\leq (1+\epsilon)(x_i)^2/n$ żołnierzy.
Uwaga Żołnierze się synchronizują, bo inaczej byłoby zbyt łatwo połączyć się z centrum.

Szkic rozwiązania Nazwijmy żołnierzy kulami a centra urnami ;) Zgodnie ze wskazówką,

$$\begin{aligned} E[x_{i+1}] &= \sum_{j=1}^{x_i} Pr[j\text{-ta kula nie jest samotna w urnie}] \\ &\leq \sum_{j=1}^{x_i}\sum_{k\neq j} Pr[j\text{-ta i } k\text{-ta kula w tej samej urnie}] \\ &= x_i(x_i-1)\frac{1}{n} < x_i^2/n, \end{aligned}$$

gdzie nierówność to union bound. Z nierówności Markowa, $P[x_{i+1} > (1+\epsilon)(x_i)^2/n] < 1/(1+\epsilon)$. Czyli (z rozkładu geometrycznego) oczekiwana liczba minut (od początku i-tej minuty) potrzebna, żeby zostało $\leq (1+\epsilon)(x_i)^2/n$ kul wynosi $\leq 1/(1-\frac{1}{1+\epsilon})$, czyli $O(1)$ dla każdego ustalonego $\epsilon > 0$.

Teraz (podobnie jak np. w problemie kolekcjonera kuponów) dzielimy czas na epoki, i-ta epoka sie konczy gdy mamy po raz pierwszy co najwyżej $y_i = (1+\epsilon)y_{i-1}^2/n$ kul. Ponieważ pokazaliśmy, że jedna epoka trwa $O(1)$, widać, że wystarczy oszacować liczbę epok.

Na pierwszy rzut oka to wygląda słabo, bo dla $y_0 = n$ ciąg y rośnie zamiast maleć. Wynika to stąd, że nasze oszacowanie jest słabe gdy y_{i-1} jest duże, czyli np. na samym początku. Policzmy więc w inny sposób, ile kul znika po pierwszej minucie (liczyliśmy to na zajęciach). Wartość oczekiwana wynosi $\sum_{i=1}^{n} Pr[i\text{-ta kula samotna}] = n(\frac{n-1}{n})^n = n(1-\frac{1}{n})^n \leq n/e$, gdzie korzystamy ze standardowej nierówności $1+x \leq e^x$. Stąd, wartość oczekiwana kul, które zostają to $(1-1/e)n$. Znowu na podstawie nierówności Markowa i rozkładu geometrycznego po $O(1)$ minutach (w sensie wartości oczekiwanej) będzie niewiele więcej, tzn. cn kul dla $c \in (1-1/e, 1)$. Jeśli podstawimy teraz $y_0 = cn$, to po $i = \Omega(\log\log n)$ krokach dostajemy $y_i = (1+\epsilon)^{2^i-1}nc^{2^i} < n\cdot(c(1+\epsilon))^{2^i} = O(1)$, jeśli tylko dobierzemy c i ϵ tak, żeby $c(1+\epsilon) < 1$.

Zadanie 3: Niezwykłe ścieżki [10 punktów] Niech $G = (V, E)$ będzie spójnym grafem nieskierowanym. Pokaż, że można przypisać krawędziom G wagi ze zbioru $\{1, \ldots, W\}$ tak, aby:

a) między każdą parą wierzchołków istniała dokładnie jedna najkrótsza ścieżka,

b) najkrótsze ścieżki między różnymi parami wierzchołków miały różne długości,

oraz $W = O(poly(n))$, $n = |V|$. Jak dużego W potrzebujesz?
Wskazówka Jeśli metoda, której probujesz użyc nie działa, to spróbuj ją nieco zmodyfikować.

Szkic rozwiązania Idea: wylosujemy wagi krawędzi i pokażemy, że z dodatnim prawdopodobieństwem zajdą pożądane zdarzenia. Skorzystamy z lematu o izolacji (patrz. np. notatki Cezarego Bartoszuka). U to oczywiście krawędzie. Przyjmijmy, że $\mathcal{F}_{x,y}$ jest rodziną wszystkich ścieżek łączących pewną parę wierzchołków x i y. Lemat zastosowany dla $\mathcal{F}_{x,y}$ mówi nam, że jeśli wagi będą losowane ze zbioru $\{1, \ldots, 2|E|\}$ to z prawdopodobieństwem $1/2$ między x i y będzie dokładnie 1 najkrótsza ścieżka (nazwijmy to zdarzenie Z_{xy}). My chcielibyśmy, żeby z niezerowym prawdopodobieństwem to zdarzenie zachodziło *równocześnie dla każdej pary*. Innymi słowy, chcemy pokazać, że prawdopodobieństwo, że zdarzenie nie zachodzi dla *pewnej pary* jest < 1, czyli że

$$P(\bigcup_{x,y \in V} \overline{Z_{xy}}) < 1.$$

Na podstawie „union bound”

$$P(\bigcup_{x,y \in V} \overline{Z_{xy}}) \leq \sum_{x,y \in V} P(\overline{Z_{xy}}). \tag{1}$$

Na podstawie naszych rozważań dostajemy zbyt słabe szacowanie (przez $\binom{|V|}{2} \cdot \frac{1}{2}$ zamiast przez 1). Byłoby dobrze, gdyby dla pojedynczej pary prawdopodobieństwo porażki było małe, np mniejsze niż $1/n^2$. Żeby zmniejszyć to prawdopodobieństwo wystarczy oczywiście losować wagi z większego zbioru. Jeśli weźmiemy zbiór $\{1, \ldots, 2|E| \cdot |V|^2\}$ i przepiszemy dowód lematu o izolacji to dostaniemy prawdopodobieństwo porażki $1/(2|V|^2)$, a stąd i z (1) $P(\bigcup_{x,y \in V} \overline{Z_{xy}}) < 1/4$.

W punkcie b) podobnie, ale teraz dla każdej czwórki wierzchołków (a, b, x, y) stosujemy lemat dla rodziny $\mathcal{F}_{a,b} \cup \mathcal{F}_{x,y}$.

Zadanie 4: Listowe słowa Thuego [10 punktów] Rozważamy słowa nad pewnym nieskończonym alfabetem Σ. Mówimy, że słowo jest bezkwadratowe gdy nie zawiera podsłowa postaci ww (gdzie w jest niepustym słowem). Dla funkcji $L : \{1, \ldots, n\} \to \binom{\Sigma}{k}$ mówimy, że słowo $x = x_1 x_2 \ldots x_n$ jest *zgodne* z L, gdy $x_i \in L(i)$ dla każdego $i = 1, \ldots, n$ (za pomocą $\binom{\Sigma}{k}$ oznaczamy wszystkie k-elementowe podzbiory zbioru Σ). Innymi słowy, $L(i)$ to dozwolone litery dla i-tego symbolu słowa x.

Pokaż, że istnieje stała k, taka że dla dowolnego n i dla dowolnej funkcji $L : \{1, \ldots, n\} \to \binom{\Sigma}{k}$ istnieje słowo bezkwadratowe długości n zgodne z L.
Wskazówki: $1 + x \leq e^x$ dla każdego x, oraz $1 - x \geq e^{-2x}$ dla $0 \leq x \leq 1/2$.

Szkic rozwiązania Stosujemy Lemat Lokalny Lovasza (wersja ogólna, nie symetryczna). „Złe” zdarzenia są postaci $A_{i,d}$ = na pozycji i pojawia się kwadrat długości $2d$. Stąd $P[A_{i,d}] = (1/k)^d$. Teraz trzeba tak dobrać wartości $X_{i,d}$, żeby

$$P[A_{i,d}] < x_{i,d} \cdot \prod_{(j,d') \in N(i,d)} (1 - x_{j,d'})$$

gdzie $N(i, d)$ oznacza zbiór takich (j, d'), że podsłowo długości $2d$ o początku w i zahacza o podsłowo długości $2d$ o początku w j. Jeśli nie interesuje nas stała k, kandydatów na $x_{i,d}$ jest wiele, widać, że powinniśmy brać funkcje, które maleją względem d (ale wolniej niż $P[A_{i,d}]$). Chyba najwygodniej wziąć e^{-d} i zadziała. P. Michał Pilipczuk podstawił $k^{-d/4}$ i też wyszło (zapewne z lepszą stałą k).