8. cvičení

Datové struktury I, 22. 11. 2024 https://iuuk.mff.cuni.cz/~chmel/2425/ds1/

Úloha 1 (Nezávislost a univerzalita)
Dokažte následující:

- pokud je systém hashovacích funkcí (k,c)-nezávislý, je také $(k-1,c)$-nezávislý (pro $k \geq 2$),
- pokud je systém hashovacích funkcí $(2,c)$-nezávislý, je též c-univerzální.

Řešení

- Chceme ukázat $(k-1,c)$-nezávislost, tedy máme dané $x_1,\dots,x_{k-1} \in \mathcal{U}, a_1,\dots,a_{k-1} \in [m]$. Zvolme si dále $x \neq x_i \forall i \in [k-1]$ (takové existuje z k-nezávislosti). Chceme $\Pr_h[h(x_1) = a_1 \wedge \ldots \wedge h(x_{k-1}) = a_{k-1}] = \sum\limits_{a\in[m]} \Pr_h[h(x_1) = a_1 \wedge \ldots \wedge h(x_{k-1}) = a_{k-1} \wedge h(x) = a] \leq \sum\limits_{a\in[m]} \frac{c}{m^k} = \frac{c}{m^{k-1}}$.
- Mějme tedy $x \neq y \in \mathcal{U}$. Chceme omezit shora $\Pr_h[h(x) = h(y)] = \sum\limits_{a\in[m]} \Pr_h[h(x) = a \wedge h(y) = a] \leq \sum\limits_{a\in[m]} \frac{c}{m^2} = \frac{c}{m}$.

Úloha 2 (Lineární systém bez konstanty)
Víme, že systém funkcí $h_{a,b}(x) = (ax + b \mod p)$ je (2,1)-nezávislý, a po modulo m se toto změní na (2,4)-nezávislost (a 2-univerzalitu). Prozkoumejme, co se stane, když nebudeme přičítat b: uvažme tedy systém funkcí $\{h_a(x) = (ax \mod p) \mod m \colon a \in \mathbb{Z}_p \setminus \{0\}\}$. Je k-univerzální pro nějaké k?
A co když dovolíme $a = 0$?

Řešení
Je 2-univerzální: $h_a(x) = h_a(y)$ odpovídá $(ax \mod p) \mod m \equiv (ay \mod p) \mod m$, tedy $((ax-ay) \mod p) \mod m \equiv 0$. To můžeme přepsat jako $a(x-y) \mod p = \ell m$ pro $l \in \{-\lfloor p/m \rfloor, \dots, \lfloor p/m \rfloor\} - \{0\}$ ($\ell \neq 0$, bo $a \neq 0$). Speciálně tedy pro každou volbu x a y máme maximálně $2\lfloor p/m \rfloor$ funkcí, ve kterých můžeme mít kolizi. Tím pádem $|\{h : h(x) = h(y)\}| \leq \frac{2p}{m}$, a pravděpodobnost je tedy $\leq \frac{2}{m}$.
Pro $a = 0$: provedeme to samé, ale ℓ může být až $2\lfloor p/m \rfloor + 1$, a tedy máme odhad $|\{h : h(x) = h(y)\}| \leq \frac{3p}{m}$ a 3-univerzalitu.

Úloha 3 (Vyloženě praktické systémy)
Uvažme systém funkcí $\mathcal{H}_1 = \{\mathrm{id}\}$, který obsahuje jedinou funkci, jež zobrazí x na x. Je $\mathcal{H}_1$ c-univerzální pro nějaké c? Je $\mathcal{H}_1$ (k,c)-nezávislý pro nějaká k a c?
Dále uvažme systém $\mathcal{H}_2 = \{h_a(x) = a \colon a \in [m]\}$. Dokažte, že tento systém je (1,1)-nezávislý. Dále ukažte, že $\mathcal{H}_2$ není $(2,c)$-nezávislý ani c-univerzální pro žádné c.

Řešení
$\mathcal{H}_1$ je ε-univerzální pro každé $\varepsilon > 0$. Problém je, že $\Pr[h(x) = x] = 1$, a tedy, pokud $|\mathcal{U}| > 1$, pak nemůže být jakkoliv nezávislá.
U druhého systému: (1,1)-nezávislost plyne z toho, že $\Pr[h_a(x) = b] = \frac{1}{m}$, protože volíme jednu konkrétní volbu a. Na druhou stranu, pro $x \neq y$ máme $\Pr[h_a(x) = b \wedge h_a(y) = b] = \frac{1}{m} > \frac{c}{m^2}$ pro jakoukoliv konstantu, a tedy nezávislost je nemožná. Pro c-univerzalitu, evidentně $\Pr[h_a(x) = h_a(y)] = 1$, a tedy c-univerzalitu také nemáme.

Úloha 4 (Modulo univerzálního systému nemusí být univerzální)
Ukažte, že pokud máme univerzální systém hashovacích funkcí $\mathcal{H}$, pak systém $\mathcal{H}'$, kde ke každé fci navíc přidáme modulo m, už nemusí být univerzální. Formálně: Dokažte, že pro každé c a $m > c$ existuje univerzum $\mathcal{U}$ a systém $\mathcal{H}$ z $\mathcal{U}$ do $\mathcal{U}$ tak, že $\mathcal{H}$ je univerzální, ale $\mathcal{H}'$ už není c-univerzální.

Řešení
Uvažme $\mathcal{H}_1 = \{\mathrm{id}\}$ z předchozí úlohy - pak $\mathcal{H}_1 \mod m$ nemůže být c-univerzální, protože dokud $m < |\mathcal{U}|$, pak prvky 1 a $m+1$ se vždycky zobrazí na tentýž prvek 1.

Věta. Tabulkové hashování je 3-nezávislé.

Úloha 5 (Tuhle větu si dokážeme)
Dokažte předcházející větu s následujícím postupem. Mějme $a, b, c \in \mathbb{Z}_2^\ell, x \neq y \neq z \neq x \in \mathbb{Z}_2^w$, a používejme tabulkové hashování s d částmi. Pak chceme ukázat, že $\Pr_{h\in\mathcal{H}}[h(x) = a \wedge h(y) = b \wedge h(z) = c] \leq \frac{1}{m^3}$.

a) Prvně si uvědomme, že pokud máme jen jednu část, a tedy jednu tabulku, tvrzení je triviální.

   Dále mějme alespoň dvě části. Protože x, y, z jsou různé, musí se (po dvou) lišit alespoň v jedné části.

b) Začneme s případem, kdy existuje část i, že x^i, y^i, z^i jsou všechny různé. Mějme jakkoliv zvolené ostatní tabulky, kromě tabulky T_i. S jakou pravděpodobností můžeme zvolit funkci pro tabulku T_i tak, že $h(x) = a, h(y) = b, h(z) = c$?

c) Jinak existují (BÚNO) části i, j takové, že $z^i = x^i \neq y^i$ a $y^j = x^j \neq z^j$. Potom máme následující soustavu rovnic, kde v_x, v_y, v_z jsou vyXORované výsledky z ostatních tabulek:

$$\begin{aligned} T_i[x^i] \oplus T_j[x^j] \oplus v_x &= a \\ T_i[y^i] \oplus T_j[y^j] \oplus v_y &= b \\ T_i[z^i] \oplus T_j[z^j] \oplus v_z &= c \end{aligned}$$

   Opět si představme, že v_x, v_y, v_z už známe. S jakou pravděpodobností budou náhodně volené tabulky T_i, T_j splňovat tuto soustavu rovnic?

d) Uvědomte si, že toto stačí.

Řešení a) Máme jednu tabulku, takže máme uniformně náhodnou funkci z $\{0,1\}^\ell$ do $\{0,1\}^w$, a ta je automaticky nezávislá.

b) Máme zafixované všechny hodnoty, a víme, že $h(x)$ musí být a, tedy speciálně $T_i(x^i) = a \oplus \bigoplus_{j=1 j\neq i}^{k} T_j(x^j)$, a to je dáno jednoznačně. Pro y, z platí totéž analogicky, a tedy máme pro volbu $T_i(x^i), T_i(y^i), T_i(z^i)$ právě jednu možnost z celkem $2^{3w} = m^3$ možností, a tedy v tomto případě máme 3-nezávislost.

c) Uvědomme si, že máme vlastně jen tři hodnoty, protože $z^i = x^i$ a $y^j = x^j$, pak naše soustava rovnic je

$$\begin{aligned} T_i[x^i] \oplus T_j[x^j] \oplus v_x &= a \\ T_i[y^i] \oplus T_j[x^j] \oplus v_y &= b \\ T_i[x^i] \oplus T_j[z^j] \oplus v_z &= c \end{aligned}$$

   Kolik má tato soustava řešení? Hodnoty v_x, v_y, v_z předpokládáme, že už jsou zafixované, tedy naše soustava rovnic se zjednoduší, zároveň označíme $W = T_i[x^i], X = T_i[y^i], Y = T_j[x^j], Z = T_j[z^j]$, a máme pak

$$\begin{aligned} W \oplus Y &= a \oplus v_x \\ X \oplus Y &= b \oplus v_y \\ W \oplus Z &= c \oplus v_z \end{aligned}$$

   Tady vidíme, že když zvolíme libovolné Z, pak W, Y, X jsou jednoznačně určené. Tedy celkem máme m možných řešení této soustavy rovnic, a máme m^4 způsobů, jak W, X, Y, Z vybrat (jakýmkoliv způsobem, i když rovnice neplatí). Tedy náhodnými volbami tuto rovnost splníme s pravděpodobností $m/m^4 = 1/m^3$.

d) Přesně tak: v každém případě tedy máme pravděpodobnost toho, že se hodnoty trefí právě $1/m^3$, což je přesně to, co po nás chce definice nezávislosti.

Úloha 6 (FKS (Fredman, Komlós, Szemerédi))
Ukážeme si konstrukci (statické) hashovací tabulky pro podmnožinu S velikosti n univerza $\mathcal{U}$, která nemá žádné kolize. Možná jste narazili na konstrukci, která potřebovala $\Omega(n^2)$ paměti (přesněji paměťových buněk). My to zvládneme s lineárním počtem paměťových buněk (za předpokladu, že můžeme mít zcela náhodnou hashovací funkci, tu dokážeme sestrojit a samplovat v konstantním čase a že si ji dokážeme pamatovat v konstantním prostoru)[1].

[1] Totéž jde udělat s rozumně univerzální funkcí, tohle je jenom pro jednoduchost.

Proces stavby tabulky bude probíhat následovně: budeme stavět dvě úrovně. V první úrovni si zcela náhodnou hashovací funkcí f rozdělíme prvky S do kyblíků $B_1, \ldots, B_n$ (a označíme si $b_i := |B_i|$). V druhé úrovni pak postavíme bezkolizní tabulku pomocí konstrukce, kdy máme pro kyblík B_i tabulku velikosti $2b_i^2$ pro b_i prvků, a zkoušíme náhodně volit vhodnou hashovací funkci, dokud nemáme žádné kolize.

První úroveň: V konstantním čase si vybereme náhodnou hashovací funkci $f : \mathcal{U} \to [n]$, a tou rozdělíme S do kyblíků. Toto opakujeme, dokud neplatí, že $\sum_{i=1}^{n} b_i^2 \leq \beta n$ pro $\beta = 4$.
Chceme ukázat, že tento krok budeme ve střední hodnotě opakovat nejvýše dvakrát. Označme jako C počet kolizí.

a) Určete $\mathbb{E}[C]$.

b) Určete C v závislosti na b_i.

c) Na základě předchozích dvou hodnot určete $\mathbb{E}[\sum_{i=1}^{n} b_i^2]$.

d) Aplikujte Markovovu nerovnost na náhodnou veličinu $X = \sum_{i=1}^{n} b_i^2$ s vhodnou hodnotou, abychom dostali požadovaný výsledek. (Taky se bude hodit střední hodnota geometrického rozdělení.)

Druhá úroveň: Ve druhé úrovni pro každé $i \in [n]$ volíme v i-tém kyblíku univerzální hashovací funkci $g_i : \mathcal{U} \to [\alpha b_i^2]$ pro $\alpha = 2$. Toto opakujeme, dokud není prostá pro prvky v kyblíku B_i.
Označme jako C_x počet kolizí klíče $x \in B_i$ na druhé úrovni.

a) Shora odhadněte $\mathbb{E}[C_x]$.

b) Použijte Markovovu nerovnost a union bound, abyste shora odhadli pravděpodobnost existence prvku s aspoň jednou kolizí.

c) Kolikrát budeme muset proces opakovat? ([Sem si vložte si svou oblíbenou poznámku o střední hodnotě geometrického rozdělení.])

Řešení
První část:

a) $\mathbb{E}_h[C] = \sum_{x \neq y \in S} \Pr[h(x) = h(y)] = \binom{n}{2} \frac{1}{n} = \frac{n-1}{2}$

b) $C = \sum_{i=1}^{n} \binom{b_i}{2} \rightsquigarrow 2C = \sum_{i=1}^{n} (b_i^2 - b_i) = \sum_{i=1}^{n} (b_i^2) - n \rightsquigarrow \sum_{i=1}^{n} b_i^2 = 2C + n.$

c) $\mathbb{E}[b_i^2] = 2\mathbb{E}[C] + n = 2n - 1$

d) $\Pr[X \geq 4n] \leq \frac{2n-1}{4n} \leq \frac{1}{2}$, a tedy ve střední hodnotě budeme potřebovat 2 pokusy, než najdeme vhodnou funkci.

Druhá část:

a) $\mathbb{E}[C_x] \leq \frac{b_i}{\alpha b_i^2} = \frac{1}{\alpha b_i}$.

b) Označíme $C' = \sum_{x \in B_i} C_x$, pak $\Pr[C' \geq 1] \leq \sum_{x \in B_i} \Pr[C_x \geq 1] \leq \sum_{x \in B_i} \frac{1}{\alpha b_i} = \frac{1}{\alpha}$.

c) Ve střední hodnotě tedy potřebujeme $\alpha = 2$ pokusy pro nalezení vhodné funkce.

Užitečné definice

Definice (c-univerzální systém fcí)**.** Systém $\mathcal{H}$ funkcí $h : \mathcal{U} \to [m]$ je c-univerzální pro $c > 0$, pokud pro všechna $x \neq y$ platí $\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq \frac{c}{m}$.
Systém $\mathcal{H}$ je univerzální, pokud je c-univerzální pro nějaké $c > 0$.

Definice (k-nezávislý systém fcí)**.** Systém $\mathcal{H}$ funkcí $h : \mathcal{U} \to [m]$ je (k, c)-nezávislý pro nějaká $k \geq 1, c > 0$, pokud $\Pr_{h \in \mathcal{H}}[h(x_1) = a_1 \wedge \ldots \wedge h(x_k) = a_k] \leq \frac{c}{m^k}$ pro libovolná $x_1, \ldots, x_k$ různá, $a_1, \ldots, a_k$ ne nutně různá.
Systém $\mathcal{H}$ je k-nezávislý, pokud je (k, c)-nezávislý pro nějakou nezávislou konstantu c.

Definice (Tabulkové hashování)**.** Představme si, že chceme zahashovat n-bitové řetízky do m-bitových řetízků, kde $n = k \cdot \ell$. Řetízek $x \in \{0,1\}^n$ pak rozložíme do k částí délky ℓ, které značíme x^i. Můžeme tedy psát $x = x^1 x^2 \dots x^k$. Pak generování naší hashovací funkce $h : \{0,1\}^n \to \{0,1\}^m$ vypadá tak, že vybereme uniformně náhodně k funkcí $T_i : \{0,1\}^\ell \to \{0,1\}^m$ (tyto reprezentujeme tabulkou, proto tabulkové hashování). Vyhodnocujeme pak $h(x) = \bigoplus_{i=1}^k T_i(x^i) = T_1(x^1) \oplus T_2(x^2) \oplus \cdots \oplus T_k(x^k)$, kde $\oplus$ značí XOR (po jednotlivých bitech).

Věta (Markovova nerovnost)**.** Buď X nezáporná náhodná veličina. Pak $\forall \varepsilon > 0$ platí $P[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon}$. Ekvivalentně pro jakékoliv $d \geq 1$, $P[X \geq d \cdot \mathbb{E}[X]] \leq \frac{1}{d}$.