Problémy řešené v očekávaném čase (hešováním)

V následujících úlohách porovnejte deterministické a pravděpodobnostní řešení. Časovou složitost pravděpodobnostního řešení určíme jako worst-case složitost bez kolizí hešovacích funkcí plus očekávaný čas strávený navíc kvůli kolizím (předpokládáme, že používáme nějaký c-univerzální systém).

Příklad 1: Pro seznam n čísel na vstupu rozhodněte, zda jsou všechna různá.

Příklad 2: Pro seznam n čísel a hodnotu k na vstupu rozhodněte zda v seznamu je dvojice čísel se součtem k.

Příklad 3: Pro seznam n čísel na vstupu rozhodněte v seznamu je dvojice čísel jejichž součet je také v seznamu.

pravděpodobnost a černá magie

Příklad 4: Předpokládejme, že náhodně zvolená hešovací funkce $h : \mathcal{U} \to 2^m$ se chová náhodně. Chápejme hodnoty hešů jako binární čísla s nezávislými bity (něco trochu slabšího platí za předpokladu silné univerzálnosti). Pokud zahešujeme n (vzájemně různých) prvků, určete následující pravděpodobnosti:

- Fixní prvek má heš začínající alespoň k nulovými bity. $[P = 1/2^k]$
- Existuje prvek s hešem začínajícím alespoň k nulovými bity (horní odhad, union-bound) $[P \leq n/2^k]$
- Všem prvkům začínají heše nejvýše k nulovými bity (horní odhad, z nezávislosti) $[P \leq (1 - 1/2^k)^n]$
- Zafixujme $n = 2^x$. Definujme $\bar{x}$ jako největší počet počátečních nul všech hešů. Určete pravděpodovnost, že $|\bar{x} - x| \leq 1$ a že $\bar{x} = x$

Stručný výpočet posledního bodu:

- $P[\bar{x} \geq x + 2] \leq n * 1/2^{x+2} = 2^x * 1/2^{x+2} = 0.25$
- $P[\bar{x} \leq x - 2] \leq (1 - 1/2^{x-2})^n = (1 - 4/2^x)^{2^x} \leq e^{-4} \approx 0.02$
- $P[|\bar{x} - x| \leq 1] \geq 1 - (0.02 + 0.25) \approx 73\%$
- $P[|\bar{x} - x| = 0] \geq 1 - (e^{-2} + 1/2) \approx 36\%$

Příklad 5: Máme stream M prvků (nelze číst opakovaně). Chceme spočítat počet různých prvků m ve streamu.

Navrhněte přesné řešení. Navrhněte přibližné řešení pokud dostupná paměť je mnohem menší než m. Určete časovou a prostorovou složitost.

Hint: Aplikací předchozího cvičení umíme v konstantím prostoru s velkou pravděpodobností odhadnout počet prvků s malou (≤ 4) multiplikativní chybou. Použitím konstantně mnoha paralelních instancí (pro velkou konstantu) umíme dosáhnout velmi malé multiplikativní chyby (0.1) s velmi vysokou pravděpodobností ($\geq 99.9\%$), ale přesná analýza už je mírně komplikovaná.

Domácí úkol 7:

Mějme na vstupu seznam n hodnot a číslo m udávající kolik různých hodnot je na vstupu ($n >> m$). Chceme v očekávaném lineárním čase $O_{\mathbf{E}}(n)$ určit počty výskytů všech hodnot.

Pro řešení použijte nějaký c-univerzální hešovací systém, není podstatné který. Podstatné jsou parametry systému, tedy především velikost oboru hodnot.

Určete očekávanou časovou složitost algoritmu pomocí očekávaného počtu kolizí hodnot hešovací funkce. Určete prostorovou složitost.