Advanced Algorithms 2024 21/10, 2024

Sample Solutions 05

Lecturer: Maximilian Probst *Teaching Assistant: Jakob Nogler*

1 Hashing with Chaining via 2-wise Independence

The time required for inserting or deleting in a linked list is determined by the length of the list. Thus, in hashing with chaining, the time for any operation involving a key x depends on length of the linked list containing x stored in $H[h(x)]$. Since the linked list in $H[h(x)]$ contains all elements $y \in S$ such that $h(y) = h(x)$, we can write the expected time of an operation as

$$\mathbb{E}[1 + |H(h(x))|] = O(1 + \sum_{x \neq y \in S} \mathbb{P}[h(y) = h(x)]).$$

We bound this last probability by

$$\begin{aligned}
\sum_{x \neq y \in S} \mathbb{P}[h(y) = h(x)] &= \sum_{x \neq y \in S} \sum_{i \in [m]} \mathbb{P}[h(y) = i \mid h(x) = i] \cdot \mathbb{P}[h(x) = i] \\
&= \sum_{x \neq y \in S} \sum_{i \in [m]} \mathbb{P}[h(y) = i] \cdot \mathbb{P}[h(x) = i] && \text{(2-wise independence)} \\
&= \sum_{x \neq y \in S} \sum_{i \in [m]} \frac{1}{m^2} = \frac{n}{m} && \text{(uniformity).}
\end{aligned}$$

2 Extending Hash Functions to Non-Prime Domains

Intuitively, we would like to apply the same reasoning as in the script. However, this approach does not extend to $g_k(x)$ in this case because $[m]$ is not necessarily a field.

To circumvent this, we define the function $h_k(x) = \sum_{i=1}^{k} a_i x^{i-1} \mod p$, omitting the final modulo m operation. Now, for any $Y_1, Y_2, \ldots, Y_k \in [U]$ where the Y_i's are pairwise distinct, i.e., $Y_i \neq Y_j$ for all $i \neq j$, we can express $h_k(Y_\ell)$ for all $1 \leq \ell \leq k$ as a system of linear equations.

$$\begin{pmatrix} Y_1^0 & Y_1^1 & \ldots & Y_1^{k-1} \\ Y_2^0 & Y_2^1 & \ldots & Y_2^{k-1} \\ \ldots & \ldots & \ldots & \ldots \\ Y_k^0 & Y_k^1 & \ldots & Y_k^{k-1} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \ldots \\ a_k \end{pmatrix} = \begin{pmatrix} h_k(Y_1) \\ h_k(Y_2) \\ \ldots \\ h_k(Y_k) \end{pmatrix}$$

Since the matrix on the left is a Vandermonde matrix and $[p]$ forms a field, there is a one-to-one correspondence between the k-tuples $(a_1, a_2, \ldots, a_k)$ and $(h_k(Y_1), h_k(Y_2), \ldots, h_k(Y_k))$. Given that $a_1, a_2, \ldots, a_k$ are uniformly distributed in $[p]^k$, it follows that $(h_k(Y_1), h_k(Y_2), \ldots, h_k(Y_k))$ is also uniformly distributed in $[p]^k$. We derive that for each $(i_1, i_2, \ldots, i_k) \in [p]^k$, we have

$$\mathbb{P}[h_k(x_1) = i_1 \wedge \cdots \wedge h_k(x_k) = i_k] = 1/p^k.$$

Thus, for each $x_1 \in [U]$ and $i_1 \in [p]$ (and any $x_2, \ldots, x_k$ such that $x_i \neq x_j$ for $i \neq j$) we have

$$\mathbb{P}[h_k(x_1) = i_1] = \sum_{(i_2, \ldots, i_k) \in [p]^{k-1}} \mathbb{P}[h_k(x_1) = i_1 \wedge \cdots \wedge h_k(x_k) = i_k] = p^{k-1} \cdot 1/p^k = 1/p.$$

This final discussion summarizes what was previously covered in the lecturespecifically, that the hash function h_k is both uniform and k-uniform. Now, let us build on this to demonstrate that similar properties apply to g_k.

To this end, set $p' \in [m]$ to be the number such that $p' = p \mod m$. Further, for an integer $x \in [m]$, let $m(x)$ denote the number of integers $y \in [p]$ such that $y = x \mod m$. Note that $m(x) = \lfloor p/m \rfloor$ if $x < p'$, and $m(x) = \lfloor p/m \rfloor + 1$ otherwise.

From this, we obtain uniformity since for each $j \in [m]$ and $x \in [U]$ we have

$$\begin{aligned}\mathbb{P}[g_k(x) = j] &= \sum_{i \in [p] \,:\, i = j \mod m} \mathbb{P}[h_k(x) = i] \\ &= p^{-1} \cdot m(j) \\ &\leq p^{-1}(\lfloor p/m \rfloor + 1) \\ &\leq p^{-1}(p/m + 1) \\ &= (1/m + 1/p) \leq (1 + 1/2c) \cdot 1/m.\end{aligned}$$

To show k-wise independence, observe that for a tuple $(j_1, j_2, \ldots, j_k) \in [m]^k$, there are $\prod_{\ell=1}^{k} m(j_\ell)$ distinct tuples $(i_1, i_2, \ldots, i_k) \in [p]^k$ such that $i_\ell = j_\ell \mod m$ for all $1 \leq \ell \leq k$. Thus, for any $(j_1, j_2, \ldots, j_k) \in [m]^k$ and $(x_1, \ldots, x_k) \in [U]^k$, we have

$$\mathbb{P}[g_k(x_1) = j_1 \wedge \cdots \wedge g_k(x_k) = j_k] = p^{-k} \cdot \prod_{\ell=1}^{k} m(j_\ell).$$

We conclude

$$\mathbb{P}[g_k(x_1) = j_1 \wedge \cdots \wedge g_k(x_k) = j_k] = p^{-k} \cdot \prod_{\ell=1}^{k} m(j_\ell) = \prod_{\ell=1}^{k} \left(m(j_\ell) \cdot p^{-1}\right) = \prod_{\ell=1}^{k} \mathbb{P}[g_k(x_\ell) = j_\ell].$$

3 Linear Probing with 3-wise Independence

We use a proof similar to the one in the script that shows how a 5-independent hash function enables linear probing with an expected time complexity of $O(1)$ per operation. That is, given an element x, we define a *run* $R(x)$ as it is defined there. Moreover, we partition $[m]$ into *dyadic intervals* and consider *ℓ-intervals*, which we say is *almost-full* if at least $2^\ell \cdot \frac{3}{4}$ items from $S \setminus \{x\}$ hash into I.

We can recycle the proof of the script up to the point where we need to bound the probability P_ℓ that a particular ℓ-interval is nearly full. Since the hash function is now 3-independent rather than 5-independent, we can no longer use the 4th moment bound. Instead, we require a corresponding lemma that applies to two random variables instead of four. To this end, it suffices to use Chebyshev's inequality.

Fix an ℓ-interval I. For each $y \in S \setminus \{x\}$, let Y_y denote the indicator random variable for whether $h(y) \in I$. Define $X = \sum_{y \in S \setminus \{x\}} Y_y$, so that

$$P_\ell = \Pr\left[|\{y \in S \setminus \{x\} \mid h(y) \in I\}| \geq \frac{3}{4} \cdot 2^\ell\right] = \Pr\left[X \geq \frac{3}{4} \cdot 2^\ell\right].$$

Recall that $\mathbb{E}[X] = \frac{n}{m} \cdot 2^\ell \leq \frac{2}{3} \cdot 2^\ell$. Now, we apply Chebyshev's inequality:

$$\Pr\left[X \geq \frac{3}{4} \cdot 2^\ell\right] = \Pr\left[X - \mathbb{E}[X] \geq \frac{1}{12} \cdot 2^\ell\right] \leq \Pr\left[|X - \mathbb{E}[X]| \geq \frac{1}{12} \cdot 2^\ell\right] \leq 144 \cdot \frac{\mathrm{Var}[X]}{2^{2\ell}}.$$

To bound the variance, we use 2-independence and the fact that for all $y \in S \setminus \{x\}$ we have $\mathrm{Var}[Y_y] \leq \mathbb{E}[Y_y]$:

$$\mathrm{Var}[X] = \mathrm{Var}\left[\sum_{y \in S\setminus\{x\}} Y_y\right] = \sum_{y \in S\setminus\{x\}} \mathrm{Var}[Y_y] \leq \sum_{y \in S\setminus\{x\}} \mathbb{E}[Y_y] \leq \frac{n}{m} \cdot 2^\ell \leq \frac{2}{3} \cdot 2^\ell.$$

Thus, we can bound the probability that I is nearly full as follows:

$$P_\ell = \Pr\left[X \geq \frac{3}{4} \cdot 2^\ell\right] \leq 144 \cdot \frac{\mathrm{Var}[X]}{2^{2\ell}} \leq \frac{96}{2^\ell}.$$

Thereby, similarly to the proof of the script, we derive that $R(x)$ is of size $[2^{\ell+2}, 2^{\ell+3})$ with probability $O(1/2^\ell)$, and we obtain a total expected runtime of

$$O(1) + \sum_{\ell=0}^{\log_2 m} O(1/2^\ell) \cdot O(2^\ell) = O(1 + \sum_{\ell=0}^{\log_2 m} 1) = O(\log n).$$

4 Method of Moments

1. If $c > k/2$, then we have at least one index j_h with multiplicity $m_h = 1$. It remains to use k-wise independence of the random variables to obtain

   $$\mathbb{E}[(Y_{j_1} - p)(Y_{j_2} - p) \cdots (Y_{j_c} - p)] = \mathbb{E}[(Y_{j_1} - p)^{m_1}]\mathbb{E}[(Y_{j_2} - p)^{m_2}] \cdots \mathbb{E}[(Y_{j_c} - p)^{m_c}]$$

   and then use that $\mathbb{E}[(X_{j_h} - p)^{m_h}] = \mathbb{E}[X_{j_h} - p] = 0$.

   For $c \leq k/2$, we use that for every h, we have either $m_h = 1$ or $(X_{j_h} - p)^{m_h} \leq (X_{j_h} - p)^2$. In the former case, as argued before, we have $\mathbb{E}[(X_{j_h} - p)^{m_h}] = 0$. In the latter case, we have

   $$\mathbb{E}[(X_{j_h} - p)^{m_h}] \leq \mathbb{E}[(X_{j_h} - p)^2] = \mathbb{E}[X_{j_h}^2] - 2\mathbb{E}[X_{j_h}]p + p^2 = \mathbb{E}[X_{j_h}] - 2p^2 + p^2 = p - p^2 \leq p.$$

   Thus, in either case $\mathbb{E}[(X_{j_h} - p)^{m_h}] \leq p$. Again the upper bound of p^c then follows by using the k-wise independence of the variables X_{j_h}.

2. Let us first upper bound the number of indices $i_1, i_2, \ldots, i_k \in [n]$ that have the same distinct indices $j_1 < j_2 < \ldots < j_c$. Since each index i_h can take at most c values, we can upper bound crudely by c^k.

   Combined with the fact that for all $c \leq k/2$ all sets of distinct indices $j_1 < j_2 < \ldots < j_c$ (no matter the multiplicities) have expected value at most p^c, we obtain

   $$\begin{aligned}
   \mathbb{E}[(X - \mu)^k] &= \sum_{i_1, i_2, \ldots, i_k \in [n]} \mathbb{E}[(Y_{i_1} - p)(Y_{i_2} - p) \cdots (Y_{i_k} - p)] \\
   &< \sum_{c=1}^{k/2} \sum_{j_1 < j_2 < \ldots < j_c \in [n]} c^k \cdot p^c \\
   &< \sum_{c=1}^{k/2} n^c \cdot c^k \cdot p^c \\
   &< 2 \cdot n^{k/2} \cdot (k/2)^k \cdot p^{k/2} \\
   &= O((k/2)^k (np)^{k/2}).
   \end{aligned}$$

   It remains to use that by Markov's inequality, we have

   $$\mathbb{P}[|X - \mu| > d\sqrt{\mu}] = \mathbb{P}[(X - \mu)^k > d^k \mu^{k/2}] \leq \frac{\mathbb{E}[(X - \mu)^k]}{d^k \mu^{k/2}}.$$

And by our calculation above, we have

$$\frac{\mathbb{E}[(X-\mu)^k]}{d^k \mu^{k/2}} = O\left(\frac{(k/2)^k \cdot (np)^{k/2}}{d^k \mu^{k/2}}\right) = O\left(\frac{(k/2)^k}{d^k}\right)$$

using $np = \mu$.