

Honors Algorithms
G22.3520-001 Fall 2007

Lecture 1

Honors Algorithms

G22.3520-001 Fall 2007

Instructor: Victor Shoup, WWH 511

Office hours: Mon

Problem review sessions: TBA

A “hybrid” course:

- Algorithms (text = CLRS)
- Automata and Complexity (text = Sipser)

Grading:

- problem sets 40%
- final exam 60%
 - doubles as CS Dept. Algorithms Exam

Course web page: <http://www.cs.nyu.edu/courses/fall07/G22.3520-001/index.html>

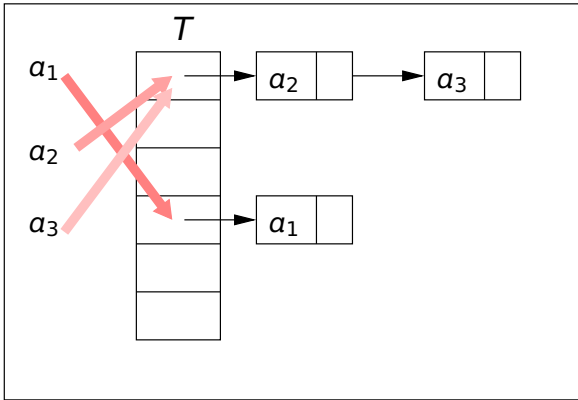
Course mailing list: See web page – *be sure to subscribe!*

Hashing

Reading: CLRS, Ch. 11, appendix C.1–4; CINTA, §§6.1–7, §6.10.

- \mathcal{U} – a set of “data items”
- $T[0 \dots m - 1]$ – a table for storing data,
*indices are called **slots**, **buckets**, or **bins***
- $h : \mathcal{U} \rightarrow \{0, \dots, m - 1\}$ – a “hash function”
maps data items to slots
- A **collision** is a pair (a, b) such that $a \neq b$ but
 $h(a) = h(b)$

Resolving collisions by chaining



Dictionary Operations:

- *insert(a)*: insert a in the linked list $T[h(a)]$
- *search(a)*: search for a in $T[h(a)]$
- *delete(a)*: search for and delete a in $T[h(a)]$

Running times:

- insert – $O(1)$
- search, delete – $O(n)$ (worst case)

Worst case occurs when all items hash to the same slot

Better: choose a *random* hash function
hopefully — no “pile ups”

Universal Hashing [Carter & Wegman, 1975]

- \mathcal{K} – a finite, non-empty set of **hash keys**
- $\mathcal{H} = \{h_k\}_{k \in \mathcal{K}}$ – a **family** of hash functions
 $h_k : \mathcal{U} \rightarrow \{0, \dots, m-1\}$, indexed by $k \in \mathcal{K}$

Def'n: \mathcal{H} is called **universal** if for all $a, b \in \mathcal{U}$ with $a \neq b$,

$$|\{k \in \mathcal{K} : h_k(a) = h_k(b)\}| \leq \frac{|\mathcal{K}|}{m}.$$

Probabilistic interpretation: if R is a random variable, uniformly distributed over \mathcal{K} , then

$$\Pr[h_R(a) = h_R(b)] \leq \frac{1}{m}.$$

Crash Course: Discrete Probability Theory

Ω : sample space (finite or countably infinite)

$\Pr : \Omega \rightarrow [0, 1]$: probability distribution

$$\sum_{\omega \in \Omega} \Pr(\omega) = 1$$

Uniform distribution: Ω finite, $\Pr(\omega) = 1/|\Omega|$

$\forall \omega \in \Omega$

$\mathcal{A} \subseteq \Omega$: event, $\Pr[\mathcal{A}] := \sum_{\omega \in \mathcal{A}} \Pr(\omega)$

Union Bound:

- $\Pr[\mathcal{A} \cup \mathcal{B}] \leq \Pr[\mathcal{A}] + \Pr[\mathcal{B}]$
- $\Pr[\mathcal{A} \cup \mathcal{B}] = \Pr[\mathcal{A}] + \Pr[\mathcal{B}] - \Pr[\mathcal{A} \cap \mathcal{B}]$

Conditional probabilities

$B \subseteq \Omega$, $\Pr[B] \neq 0$

$$\Pr(\omega | B) := \begin{cases} \frac{\Pr(\omega)}{\Pr[B]} & \text{if } \omega \in B \\ 0 & \text{otherwise} \end{cases}$$

$\Pr(\cdot | B)$: conditional distribution given B

$$\Pr[\mathcal{A} | B] := \sum_{\omega \in \mathcal{A}} \Pr(\omega | B) = \Pr[\mathcal{A} \cap B] / \Pr[B]$$

\mathcal{A} and B are **independent** if

$$\Pr[\mathcal{A} \cap B] = \Pr[\mathcal{A}] \Pr[B]$$

or equivalently, if

$$\Pr[\mathcal{A} | B] = \Pr[\mathcal{A}]$$

Total probability

Suppose $\{\mathcal{B}_i\}_{i \in I}$ is a partition of Ω

Then

$$\begin{aligned}\Pr[\mathcal{A}] &= \sum_{i \in I} \Pr[\mathcal{A} \cap \mathcal{B}_i] \\ &= \sum_{i \in I} \Pr[\mathcal{A} \mid \mathcal{B}_i] \Pr[\mathcal{B}_i]\end{aligned}$$

Random variables

Suppose S is a set

A function $X : \Omega \rightarrow S$ is called a **random variable**

For $s \in S$, " $X = s$ " is the event $\{\omega \in \Omega : X(\omega) = s\}$

Let $Y : \Omega \rightarrow T$ be another random variable

X and Y are **independent** if the events $X = s$ and $Y = t$ are independent for all $s \in S$ and $t \in T$

Indicator variables: if \mathcal{A} is an event, the corresponding indicator variable is X , where

$$X(\omega) := \begin{cases} 1 & \text{if } \omega \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}$$

Distribution of a random variable

Let X be a random variable whose image is S

The distribution of X : $\Pr_X : S \rightarrow [0, 1]$, where
 $\Pr_X(s) := \Pr[X = s]$, $\forall s \in S$

If \Pr_X is the uniform distribution on S , then we say
 X is uniformly distributed over S

Expectation

Let X be a random variable with image $S \subseteq \mathbb{R}$

Expectation:

$$E[X] := \sum_{\omega \in \Omega} X(\omega) \Pr(\omega) = \sum_{s \in S} s \Pr[X = s]$$

Linearity of expectation: $E[X + Y] = E[X] + E[Y]$

Products: if X and Y are independent, then $E[XY] = E[X]E[Y]$

A useful fact: If X takes non-negative integer values, then $E[X] = \sum_{i \geq 1} \Pr[X \geq i]$

$$\text{Proof: } p_i := \Pr[X = i], \quad \begin{array}{ccc} p_1 & & \\ p_2 & p_2 & \\ p_3 & p_3 & p_3 \\ \dots & \dots & \dots \end{array}$$

Expectation of indicator variables

If X is the indicator for an event \mathcal{A} , then

$$E[X] = \Pr[\mathcal{A}]$$

proof: $E[X] = 1 \cdot \Pr[X = 1] + 0 \cdot \Pr[X = 0] = \Pr[\mathcal{A}]$

Some inequalities

Markov's inequality: Suppose X takes only non-negative values. Then for every $\alpha > 0$, we have $\Pr[X \geq \alpha] \leq E[X]/\alpha$

Jensen's inequality (special case): for any real-valued random variable X , $E[X^2] \geq E[X]^2$

End of Crash Course!

Using Universal Hash Functions

Assume distinct items a_1, \dots, a_n are stored in table

Let $\alpha := n/m = \text{“load factor”}$

Assume R is uniformly distributed over \mathcal{K}

For $i = 1, \dots, n$, define

$$S_i := \# \text{ of items in slot } h_R(a_i)$$

That is, S_i is the number of items in the slot occupied by a_i

The values R, S_1, \dots, S_n are *random variables*.

For each $i = 1, \dots, n$, we wish to bound $E[S_i]$.

Claim: $E[S_i] \leq \alpha + 1$ for each $i = 1, \dots, n$.

Proof: for $i, j = 1, \dots, n$, define indicator variables

$$C_{ij} := \begin{cases} 1 & \text{if } h_R(a_i) = h_R(a_j) \\ 0 & \text{otherwise} \end{cases}$$

For all i, j :

$$E[C_{ij}] = \Pr[h_R(a_i) = h_R(a_j)] \begin{cases} \leq 1/m & \text{if } i \neq j \\ = 1 & \text{if } i = j \end{cases}$$

Write S_i as sum of indicator variables: $S_i = \sum_{j=1}^n C_{ij}$

By linearity of expectation:

$$\begin{aligned} E[S_i] &= \sum_{j=1}^n E[C_{ij}] = E[C_{ii}] + \sum_{j \neq i} E[C_{ij}] \\ &\leq 1 + (n-1)/m \\ &\leq \alpha + 1 \quad \text{QED} \end{aligned}$$

interpretation:

- for each i , the expected # of items in a_i 's slot (including a_i itself) is $\leq \alpha + 1$
- the expected time to perform a *single* dictionary operation is $O(\alpha + 1)$
- by linearity of expectation, expected time to perform k dictionary operations is $O(k(\alpha + 1))$

special case: $\alpha = O(1)$ (i.e., $n = O(m)$)

- expected time per operation is $O(1)$

Maximum Load: another performance measure

Suppose hash table contains items a_1, \dots, a_n , and that R is uniform over \mathcal{K}

For $s = 0, \dots, m - 1$, define

$$L_s := \# \text{ of } a_i\text{'s that hash to slot } s \text{ under } h_R$$

Set $M := \max\{L_s : s = 0, \dots, m - 1\}$

We want to bound $E[M]$, assuming universal hashing

Fact: $E[M]^2 \leq E[M^2]$

Fact: $M^2 \leq V := \sum_{s=0}^{m-1} L_s^2$

Claim: $E[V] \leq n^2/m + n$

Proof of claim: Define indicator variables

$$I_{i,s} := \begin{cases} 1 & \text{if } h_R(a_i) = s \\ 0 & \text{otherwise} \end{cases}$$

We have

$$\begin{aligned} V &= \sum_{s=0}^{m-1} L_s^2 = \sum_{s=0}^{m-1} \left(\sum_{i=1}^n I_{i,s} \right)^2 \\ &= \sum_s \left(\sum_i I_{i,s} \right) \left(\sum_j I_{j,s} \right) \\ &= \sum_{i,j} \sum_s I_{i,s} I_{j,s} = \sum_{i,j} C_{ij} \end{aligned}$$

So we have

$$V = \sum_{i,j} C_{ij}$$

and by linearity of expectation, we have

$$\begin{aligned} E[V] &= \sum_{i,j} E[C_{ij}] \\ &= \sum_i E[C_{ii}] + \sum_{i \neq j} E[C_{ij}] \\ &\leq n + n(n-1)/m \\ &\leq n^2/m + n \end{aligned}$$

QED

Corollary: $E[M] \leq \sqrt{n^2/m + n}$

Special case: $\alpha = O(1)$ (i.e., $n = O(m)$)

$$E[M] = O(\sqrt{m})$$

- This bound is tight
- Counter-intuitive: it may be the case that $E[L_s] = O(1)$ for each slot s

General fact: expected value of max may be much larger than max of expected values