Recitation 4: Bucket Sort, Applications of Median

Outline

• Applications of Median
  – Median Partitioned Quicksort
• Bucket Sort

Median Partitioned Quicksort

• Quicksort has average running time $\Theta(n \lg n)$ but a worst case running time $\Theta(n^2)$.
• Can we improve the worst case?
  – partition around the median!!
  • Finding the median runs in $\Theta(n)$ worst case!
  – Recursively sort both halves.
• $T(n) = 2T(n/2) + \Theta(n)$
  = $\Theta(n \lg n)$ by Master Theorem case 2.
• Theoretically interesting but not used much in practice - constants in $\Theta(n)$ for finding median is large.

Bucket Sort

• Linear time sorts:
  – Counting sort : Bound on the size of the numbers known
  – Radix sort : Bound on the number of bits known
• Linear time sorting of real number
  – Bucket sort: Expected run time is $\Theta(n)$ when the input is generated independently by a distribution over $[0,1)$.

Bucket Sort

• Partition $[0,1)$ into $n$ equal-sized intervals (buckets).
• Distribute the inputs into the buckets.
  – A linked list normally used
• Sort each bucket using insertion sort
• Go through the buckets in order, listing the elements.

BUCKET-SORT(A)

1. $n \leftarrow \text{length}[A]
2. \text{for } i \leftarrow 1 \text{ to } n
3. \quad \text{do insert } A[i] \text{ into } B[i\text{[nA[i]]]}
4. \text{for } i \leftarrow 0 \text{ to } n - 1
5. \quad \text{do sort list } B[i] \text{ with insertion sort}
6. \text{concatenate the lists } B[0], B[1], \ldots, B[n - 1] \text{ together in order}
Running time

- Worst case running time
  - All elements fall into the same bucket
  - $O(n^2)$
- Average case running time
  - $\Theta(n)$

Analysis

$$T(n) = \Theta(n) + \sum_{i=1}^{n} O(n^2)$$

$$E[T(n)] = E[\Theta(n)] + \sum_{i=1}^{n} E[O(n^2)]$$

(linearity of expectation)

$$= \Theta(n) + \sum_{i=1}^{n} O(n^2)$$

- We claim that $E[n_i^2] = 2 - \frac{1}{n}$
- So $E[T(n)] = \Theta(n)$

• Let

$$X_{ij} = \begin{cases} 1 & \text{if } A[j] \text{ falls in bucket } i \\ 0 & \text{otherwise} \end{cases}$$

for $i = 0, 1, \ldots, n-1$ and $j = 1, 2, \ldots, n$ (indicator random variable).

- We have

$$R_i = \sum_{j=1}^{n} X_{ij}$$

$$E[X_{ij}^2] = \frac{1}{n} + 0 \cdot \left(1 - \frac{1}{n}\right) = \frac{1}{n}$$

- When $k \neq j$, $X_{ij}$ and $X_{ik}$ are independent

$$E[X_{ij}X_{ik}] = E[X_{ij}]E[X_{ik}]$$

$$= \frac{1}{n} \cdot \frac{1}{n} = \frac{1}{n^2}$$

$$E[n_i^2] = \sum_{j=1}^{n} \frac{1}{n} + \sum_{k \neq j} \frac{1}{n}$$

$$= \frac{n-1}{n} + n(n-1) \cdot \frac{1}{n}$$

$$= 1 + \frac{n-1}{n}$$

$$= 2 - \frac{1}{n}$$

• Even if input distribution is not uniform, expected running time is still linear if $E[n_i^2] = O(1)$. 
Exercise

A ship arrives at a port, and the 40 sailors on board go ashore for revelry. When they return, each chooses a random cabin in their state of drunkenness. What is the expected number of sailors in their own cabins?

Summary

• Application of median:
  – Good worst case theoretical bound for median partitioned quicksort
• Bucket sort:
  – Linear time sorting of real number
  – Only for expected run time
  – Only for certain distributions of inputs