Bucket Sort

Bucket sort runs in linear time on the average. It assumes that the input is generated by a random process that distributes elements uniformly over the interval [0, 1).

The idea of Bucket sort is to divide the interval [0, 1) into n equal-sized subintervals, or buckets, and then distribute the n input numbers into the buckets. Since the inputs are uniformly distributed over (0, 1), we don't expect many numbers to fall into each bucket. To produce the output, simply sort the numbers in each bucket and then go through the bucket in order, listing the elements in each.

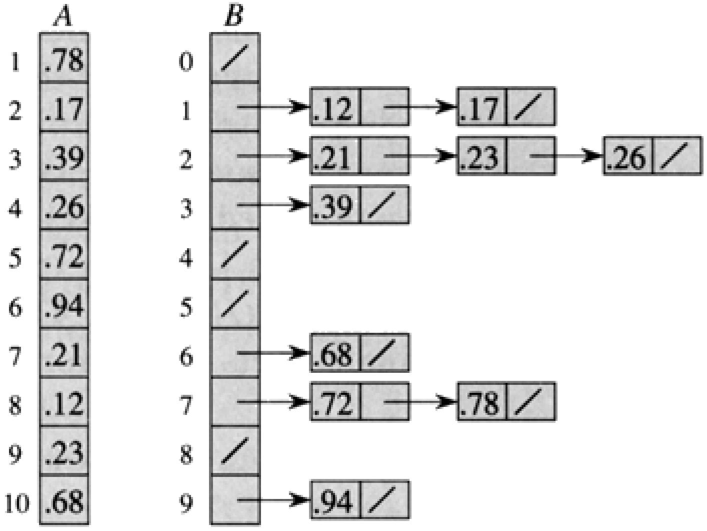
The code assumes that input is in n-element array A and each element in A satisfies $0 \le A[i] \le 1$. We also need an auxiliary array $B[0\ .\ .\ n\ -1]$ for linked-lists (buckets).

BUCKET_SORT (A)

```
1. $n \leftarrow$ length $[A]$
2. For $i = 1$ to $n$ do
3.     Insert $A[i]$ into list $B[nA[i]]$
4. For $i = 0$ to $n$-1 do
5.     Sort list $B$ with Insertion sort
6. Concatenate the lists $B[0], B[1], . . B[n-1]$
   together in order.
```

Example

Given input array A[1..10]. The array B[0..9] of sorted lists or buckets after line 5. Bucket i holds values in the interval $[i/10, (i+1)/10]$. The sorted output consists of a concatenation in order of the lists first B[0] then B[1] then B[2] ... and the last one is B[9].



Analysis

All lines except line 5 take $\mathrm{O}(n)$ time in the worst case. We can see inspection that total time to examine all buckets in line 5 is $\mathrm{O}(n\text{-}1)$ i.e., $\mathrm{O}(n)$.

The only interesting part of the analysis is the time taken by Insertion sort in line 5. Let n_i be the random variable denoting the number of elements in the bucket $B[i]$. Since the expected time to sort by INSERTION_SORT is $\mathrm{O}(n^2)$, the expected time to sort the elements in bucket $\mathrm{B}[i]$ is

$$E[\mathrm{O}({}^2n_i)] = \mathrm{O}(E[{}^2n_i]]$$

Therefore, the total expected time to sort all elements in all buckets is

$${}^{\mathrm{n-1}}\sum_{\mathrm{i=0}} O(E[{}^2n_i]) = O\,{}^{\mathrm{n-1}}\sum_{\mathrm{i=0}} (E[{}^2n_i]) \qquad \text{------------ A}$$

In order to evaluate this summation, we must determine the distribution of each random variable n

We have n elements and n buckets. The probability that a given element falls in a bucket $B[i]$ is $1/n$ i.e., Probability $= p = 1/n$. (Note that this problem is the same as that of "Balls-and-Bin" problem).

Therefore, the probability follows the binomial distribution, which has

mean: $\mathrm{E}[n_i] = np = 1$

variance: $\mathrm{Var}[n_i] = np(1- p) = 1- 1/n$

For any random variable, we have

$$
\begin{aligned}
E[^2 n_i] &= \mathrm{Var}[n_i] + E^2[n_i] \\
&= 1 - 1/n + 1^2 \\
&= 2 - 1/n \\
&= \Theta(1)
\end{aligned}
$$

Putting this value in equation A above, (do some tweaking) and we have a expected time for INSERTION_SORT, $\mathrm{O}(n)$.

Now back to our original problem

In the above Bucket sort algorithm, we observe

T(n) = [time to insert *n* elements in array *A*] + [time to go through auxiliary array *B*[0 . . *n*-1] * (Sort by INSERTION_SORT)

$$
\begin{aligned}
&= \mathrm{O}(n) + (n\text{-}1) □(n) \\
&= \mathrm{O}(n)
\end{aligned}
$$

Therefore, the entire Bucket sort algorithm runs in linear expected time.

Back