CSC473 Assignment 2: Constant Time Approximate Median

Jeff Blair: 1002177057
jeffrey.blair@mail.utoronto.ca

Jaryd Hunter: 1002725893
jaryd.hunter@mail.utoronto.ca

March 2020

Introduction:

In this paper we will discuss a way to compute the approximate median of an array with n distinct integers in constant time, with approximation factor ϵ. The naive solution of sorting the array and picking the middle element has time complexity $O(n \log n)$.

Rank Distribution:

Consider k indices $\{i_1, \ldots, i_k\}$ sampled randomly and uniformly with replacement from $[n]$. Given an array A of size n, let X be the number of integers in the set $\{A[i_1], \ldots, A[i_k]\}$ that have rank less than or equal to $\left(\frac{1}{2} - \epsilon\right) n$. We see that $P(\text{rank}(A[i_j]) \leq \left(\frac{1}{2} - \epsilon\right) n) = \frac{\left(\frac{1}{2} - \epsilon\right) n}{n} = \frac{1}{2} - \epsilon$.

$$\text{Then } X \sim \text{Binomial}(p = \tfrac{1}{2} - \epsilon, n = k)$$

Algorithm:

```python
import numpy as np
import numpy.random as npr

def ApxMedian(A,k):
    """
    A = np.array of size n
    """
    I = npr.randint(low=0, high=A.size-1, size=k)
    sample = A[I] # Returns an array of size k
    return np.sort(sample)[k//2] # O(klog k) is constant with respect to n
```

Analysis:

Let Z be the returned value from ApxMedian. We will show that Z has rank between $\left(\frac{1}{2} - \epsilon\right) n + 1$ and $\left(\frac{1}{2} + \epsilon\right)$ in A.

Let L be the random variable of how many elements in the sample have rank less than or equal to $(\frac{1}{2} - \epsilon)n$, and let H be the random variable of how many elements in the sample have rank greater than or equal to $(\frac{1}{2} + \epsilon)n + 1$. L Follows the same binomial distribution as X, so $E[L] = k(\frac{1}{2} - \epsilon)$, and $V[L] = k(\frac{1}{2} - \epsilon)(\frac{1}{2} + \epsilon)$. H also follows a binomial distribution with

$$p = \frac{n - (\frac{1}{2} + \epsilon)n}{n} = 1 - \frac{1}{2} - \epsilon = \frac{1}{2} - \epsilon$$

Then, $E[H] = k(\frac{1}{2} - \epsilon)$ and $V[H] = k(\frac{1}{2} - \epsilon)(\frac{1}{2} + \epsilon)$.

Note,

$$
\begin{aligned}
P[L > \frac{k}{2}] &= P[L - E[L] > \frac{k}{2} - E[L]] \\
&< \frac{k(\frac{1}{2} - \epsilon)(\frac{1}{2} + \epsilon)}{(\frac{k}{2} - k(\frac{1}{2} - \epsilon))^2} && \text{Using Chebyshev inequality} \\
&= \frac{k(\frac{1}{2} - \epsilon)(\frac{1}{2} + \epsilon)}{k^2\epsilon^2} \\
&= \frac{\frac{1}{4} - \epsilon^2}{k\epsilon^2} \\
&\leq \frac{\frac{1}{4} - \epsilon^2}{2} && \text{Let, } k \geq \frac{2}{\epsilon^2} \\
&= \frac{1}{8} - \frac{\epsilon^2}{2} \\
&\leq \frac{1}{8}
\end{aligned}
$$

Since, $E[H] = E[L]$ and $V[H] = V[L]$ the same result applies to $P[H > \frac{k}{2}]$.

$$
\begin{aligned}
P\left[(\frac{1}{2} - \epsilon)n + 1 \leq \text{rank}(Z) \leq (\frac{1}{2} + \epsilon)n\right] &= P\left[\text{rank}(Z) \geq ((\frac{1}{2} - \epsilon)n) + 1\right] - P\left[\text{rank}(Z) > (\frac{1}{2} + \epsilon)n\right] \\
&= 1 - P[L > \frac{k}{2}] - P[H > \frac{k}{2}] \\
&\geq 1 - \frac{1}{8} - \frac{1}{8} \\
&\geq \frac{3}{4}
\end{aligned}
$$

So with probability greater than 75%, ApxMedian returns an approximate median, with rank error bounded by $\max(\epsilon n, |1 - \epsilon n|)$