Randomized Selection

Restate the **Find** algorithm described in Section 1.4 (Page 15) of the book [MR95] as follows:

**Input:**
1. a set \( X \) of \( n \) distinct numbers, and
2. a positive integer \( k \) that is no more than \( n \).

**Output:**
the \( k \)-th smallest number in \( X \).

**Algorithm:**

```plaintext
algorithm RandSelect(X, k) {
    select an \( x \) uniformly at random from \( X \);
    partition \( X \) into \( Y \) and \( Z \) by \( x \) as in RandQS;
    if \( |Y| = k - 1 \), then print \( x \);
    if \( |Y| < k - 1 \), then call RandSelect(Z, k - 1 - |Y|);
    if \( |Y| > k - 1 \), then call RandSelect(Y, k);
}
```

Show that the expected running time of **RandSelect** is \( O(n) \).

1 Properties and Ideas

The following two hints are all from [http://www.drywettex.com/minority/?u=732&a=board&bid=27751#29853](http://www.drywettex.com/minority/?u=732&a=board&bid=27751#29853).

1.1 Hint One: Reduced Factor Evaluation

Since \( k \) is a positive integer that is no more than \( n \), it is convenient to introduce a new parameter \( \epsilon = k/n \in (0, 1] \).

For specification, we name **RandSelect**(\( X, k \)), the action that the input set \( X \) is used for the argument of **RandSelect**, as “the 1st recursion.” After \( X \)
splitting into $Y$ and $Z$, we also name the action $\text{RandSelect}(Y, k)$ (or its exclusive action $\text{RandSelect}(Z, k - 1 - |Y|)$) as “the 2nd recursion.”

In the case $|Y| < k - 1$, we observe that all the probable cardinal numbers of $Z$ are \{n − 1, n − 2, . . . , n − k + 1\}. For convenience’ sake, let $Z_i$ be the event that “the cardinal number of $Z$ is $n − i$” (for $i \in \{k−1\}$), the expected cardinality of $Z$ in the 2nd recursion is

$$|Z| = \sum_{i=1}^{k-1} (\Pr[Z_i] \cdot |Z_i|) = \frac{1}{k-1} \left( (k-1)n - \frac{k(k-1)}{2} \right) = n - \frac{k}{2}. \quad (1.1.a)$$

Similarly, in the case $|Y| > k - 1$, all the probable cardinal numbers of $Y$ are \{k, k + 1, . . . , n − 1\}. Again, let $Y_i$ be the event that “the cardinal number of $Y$ is $i$” (for $i \in \{k, k+1, \ldots, n−1\}$), the expected cardinality of $Y$ in the 2nd recursion is

$$|Y| = \sum_{i=k}^{n-1} (\Pr[Y_i] \cdot |Y_i|) = \frac{1}{k-1} \left( \frac{(n-k)(n-1+k)}{2} \right) = \frac{n}{2} + \frac{k-1}{2}. \quad (1.1.b)$$

After evaluating $|Z|$ and $|Y|$, the expected cardinality used for $\text{RandSelect}$ in the 2nd recursion can be also evaluated as follows:

$$\Pr[|Y| < k - 1] \cdot |Z| + \Pr[|Y| > k - 1] \cdot |Y| = \frac{k-1}{n} \cdot [1.1.a] + \frac{n-k}{n} \cdot [1.1.b]$$

$$= \frac{k-1}{n} \cdot \left( n - \frac{k}{2} \right) + \frac{n-k}{n} \cdot \left( \frac{n}{2} + \frac{k-1}{2} \right)$$

$$= \frac{(n^2 + 2kn - 2k^2) - (3n - 2k)}{2n}$$

$$= \frac{(1 + 2\epsilon - 2\epsilon^2)n^2 - (3 - 2\epsilon)n}{2n}, \quad (1.1.1)$$

and the ratio of the expected cardinality in the 2nd recursion to that in the 1st recursion is

$$\delta = \frac{1.1.1}{n} = \frac{(1 + 2\epsilon - 2\epsilon^2)n^2 - (3 - 2\epsilon)n}{2n^2} \leq \frac{(1 + 2\epsilon - 2\epsilon^2) \cdot 2}{2} \leq \frac{3}{4} \quad (1.1.2)$$

because $1 + 2\epsilon - 2\epsilon^2 = 3/2 - 2(\epsilon - 1/2)^2$ is maximal when $\epsilon = 1/2$. 

2
Heretofore, we find that the expected cardinality used for \textbf{RandSelect} in the \textit{j-th recursion} is asymptotically $\delta^j n$. Let $m$ be the expected number of recursions necessary for returning the solution ($m$ satisfies $\delta^m n = 1$), then the expected running time of \textbf{RandSelect} is

$$\sum_{j=0}^{m} \delta^j n = \left( \sum_{j=0}^{m} \delta^j \right) n \leq 4n = O(n) \quad (1.1)$$

because $\sum_{i=0}^{m} \delta^i \leq \sum_{i=0}^{\infty} \delta^i \leq \sum_{i=0}^{\infty} (3/4)^i = 1/(1 - 3/4) = 4$.

### 1.2 Hint Two: Exact Formula

Let the expected running time of \textbf{RandSelect} be $T(n, k)$, we can obtain the following recurrence relation for $T(\cdot, \cdot)$:

$$T(1, 1) = 0; \\
T(n, k) = (n - 1) + \frac{1}{n} \left( \sum_{i=1}^{k-1} T(n - i, k - i) + \sum_{i=k+1}^{n} T(i - 1, k) \right). \quad (1.2.x)$$

The goal of this sub-session is to find the exact form of $T(n, k)$.

#### 1.2.1 Extremal Cases

- When $k = 1$, the recurrence relation (1.2.x) becomes

$$T(n, 1) = (n - 1) + \frac{1}{n} \sum_{i=2}^{n} T(i - 1, 1). \quad (1.2.1.a)$$

Replace $n$ by $n - 1$ in (1.2.1.a):

$$T(n - 1, 1) = (n - 2) + \frac{1}{n - 1} \sum_{i=2}^{n-1} T(i - 1, 1). \quad (1.2.1.b)$$

Thus,

$$nT(n, 1) - (n - 1)T(n - 1, 1) = 2(n - 1) + T(n - 1, 1)$$

$$\iff T(n, 1) - T(n - 1, 1) = \frac{2n - 1}{n}$$

$$\iff T(n, 1) = T(n - 1, 1) + 2(1 - \frac{1}{n}) = T(1, 1) + 2 \sum_{i=1}^{n} (1 - \frac{1}{i})$$

$$= 2(n - H_n) \quad (1.2.1.A)$$

*Actually, $m \geq (\lg n)/(2 - \lg 3) = \Omega(\lg n)$. 
where
\[ H_n \triangleq \sum_{i=1}^{n} \frac{1}{i} = \ln n + \gamma + O\left(\frac{1}{n}\right) \] (for \( \gamma \approx 0.5772 \)).

- When \( k = n \), the recurrence relation (1.2.x) becomes
\[ T(n, n) = (n - 1) + \frac{1}{n} \sum_{i=1}^{n-1} T(n - i, n - i). \] (1.2.1.c)

Replace \( n \) by \( n - 1 \) in (1.2.1.c) in similar way:
\[ T(n - 1, n - 1) = (n - 2) + \frac{1}{n-1} \sum_{i=1}^{n-2} T(n - i, n - i). \] (1.2.1.d)

Fortunately,
\[ nT(n, n) - (n - 1)T(n - 1, n - 1) = 2(n - 1) + T(n - 1, n - 1) \]
\[ \iff T(n, n) - T(n - 1, n - 1) = T(1, 1) + 2 \sum_{i=1}^{n} \frac{1}{i} \]
\[ \iff T(n, n) = T(n - 1, n - 1) + 2 - \frac{2}{n} = T(1, 1) + 2 \sum_{i=1}^{n} \frac{1}{i} \]
\[ = 2(n - H_n) = T(n, 1) \] (1.2.1.C)

1.2.2 General Case

For simplifying (1.2.x), we should do some phase of reductions.
1. (a) \( T(n, k) = \)
\[ (n - 1) + \frac{1}{n} \left( \sum_{i=1}^{k-1} T(n - i, k - i) + \sum_{i=k+1}^{n} T(i - 1, k) \right). \] (1.2.a)

(b) \( T(n - 1, k) = \)
\[ (n - 2) + \frac{1}{n-1} \left( \sum_{i=1}^{k-1} T(n - 1 - i, k - i) + \sum_{i=k+1}^{n-1} T(i - 1, k) \right). \] (1.2.b)

\( ^{\dagger} \)See (http://mathworld.wolfram.com/HarmonicNumber.html).
(c) \( nT(n, k) - (n - 1)T(n - 1, k) = n(1.2.a) - (n - 1)(1.2.b) \)
\[
= 2(n - 1) + T(n - 1, k) \\
+ \sum_{i=1}^{k-1} (T(n - i, k - i) - T(n - 1 - i, k - i)), \quad (1.2.c)
\]

2. (A) We can rearrange (1.2.c) to the following equation:
\[
n(T(n, k) - T(n - 1, k)) \\
= 2(n - 1) + \sum_{i=1}^{k-1} (T(n - i, k - i) - T(n - 1 - i, k - i)). \quad (1.2.A)
\]

(B) Similarly, substitute \( n - 1 \) for \( n \) in (1.2.A), we have
\[
(n - 1)(T(n - 1, k - 1) - T(n - 2, k - 1)) \\
= 2(n - 2) + \sum_{i=1}^{k-2} (T(n - 1 - i, k - 1 - i) - T(n - 2 - i, k - 1 - i)). \quad (1.2.B)
\]

(C) \( (1.2.A)-(1.2.B) \)
\[
=n(T(n, k) - T(n - 1, k)) - (n - 1)(T(n - 1, k - 1) - T(n - 2, k - 1)) \\
= 2 + T(n - 1, k - 1) + T(n - 2, k - 1). \quad (1.2.C)
\]

3. (a) The equation (1.2.c) can be also restated as the following one:
\[
T(n, k) - T(n - 1, k - 1) = \frac{2}{n} + T(n - 1, k) - T(n - 2, k - 1) \\
= \left( \sum_{i=k+1}^{n} \frac{2}{i} \right) + T(k, k) - T(k - 1, k - 1) = \left( \sum_{i=k+1}^{n} \frac{2}{i} \right) + \frac{2(k - 1)}{k} \quad (1.2.1)
\]
(b) We can also rewrite (1.2.1) as
\[
T(n, k) = T(n - 1, k - 1) + \left( \sum_{i=k+1}^{n} \frac{2}{i} \right) + \frac{2(k-1)}{k}
\]
\[
= T(n - 1, k - 1) + 2 \left( H_n - H_k + 1 - \frac{1}{k} \right)
\]
\[
= T(n - k + 1, 1) + 2 \left( \sum_{i=n-k+2}^{n} H_i - \sum_{i=2}^{k} H_i + (k-1) + (H_k - 1) \right)
\]
\[
= 2 \left( (n - k + 1) - H_{n-k+1} \right) + 2 \left( \sum_{i=n-k+2}^{n} H_i - \sum_{i=2}^{k} H_i + (k-1) + (H_k - 1) \right)
\]
\[
= 2((n - k + 1) - H_{n-k+1}) + 2 \left( \sum_{i=n-k+2}^{n} H_i - \sum_{i=2}^{k} H_i + (k-1) + (H_k - 1) \right)
\]
\[
= 2((n - k + 1) - H_{n-k+1}) + 2 \left( \sum_{i=n-k+2}^{n} H_i - \sum_{i=2}^{k} H_i + (k-1) + (H_k - 1) \right)
\]
by applying the fact that
\[
\sum_{i=1}^{n-1} H_i = n(H_n - 1).
\]

4. Verify (1.2.2) by an extremal case: \( n = 2k - 1 \) is odd:
\[
(1.2.2) = 2 \left( 2k + 2 + 2kH_{2k-1} - 2(k+2)H_k \right)
\]
\[
= 2 \left( 2k(1 + H_{2k-1} - H_k) - 4H_k + 2 \right)
\]
\[
= 2 \left( 2k(1 + \ln(2 - k^{-1})) - 4 \ln k - 4\gamma + 2 - O(k^{-1}) \right)
\]
\[
\leq 4(1 + \ln 2)k = 2(1 + \ln 2)(n + 1) = O(n)
\]

Finally, the expected number of running time of \texttt{RandSelect} is exactly (1.2.2).

References