Randomized Selection

We use random sampling to select the kth smallest element of an ordereded set S .

Some definitions:

- $r_s(t)$ is the rank of an element t in set S.
- $S_{(i)}$ is the ith smallest element of S.

We sample with replacement, meaning that we can chose the same element several times.

LazySelect

Input: Ordered set S of n elements and an integer $k \leq n$. **output:** kth smallest element of S.

1. $x = kn^{-1/4}$, $\ell = \max\{\lfloor x - \sqrt{n} \rfloor, 1\}$, and $h = \min\{\lceil x + \sqrt{n} \rceil, n^{3/4}\}$.
2. Pick $n^{3/4}$ elements form S, chosen i.u.r. with replacement. Call this set R.
3. Sort R in time $O(n^{3/4} \log n) = O(n)$.
4. Let $a = R_{(\ell)}$ and $b = R_{(h)}$. Compare a and b to every element of S and compute $r_S(a)$ and $r_S(b)$.
5. Now compute a subset P
   - If $k < n^{1/4}$ then $P = \{y \in S \mid y \leq b\}$,
   - else If $k > n - n^{1/4}$, let $P = \{y \in S \mid y \geq b\}$,
   - else If $k \in [n^{1/4}, n - n^{1/4}]$, let $P = \{y \in S \mid a \leq y \leq b\}$.

   Check whether $S_{(k)} \in P$ and $|P| \leq 4n^{3/4} + 2$. If not, repeat steps 1-4 until such P is found.
6. By sorting P in $O(|P| \log |P|)$ steps, identify $P_{k - r_s(a) + 1}$, which is $S_{(k)}$.

Analysis of LazySelect

The idea of the algorithm is to identify two elements a and b such that both of the following statements hold with high probability $(1 - 1/n^\alpha)$:

- The element $S_{(k)}$ that we seek is in P.
- The set P of elements between a and b is not very large, so that we can sort it in time $O(n)$.

Theorem With probability $1 - O(n^{-1/4})$, LazySelect finds $S_{(k)}$ on the first pass and thus performs only $2n + o(n)$ comparisions.

We have to consider three cases, here we consider the case $k \in [n^{1/4}, n - n^{1/4}$ and $P = \{y \in S \mid a \leq y \leq b\}$. The alnalysis of the other two cases is similar.

Case 1 We fail 1) if $a > S_{(k)}$ or $b < S_{(k)}$. This means fewer than ℓ samples should be smaller than $S_{(k)}$/ at least h samples should be smaller than $S_{(k)}$.

Let's consider the event $a > S_{(k)}$. Let $X_i = 1$ if the ith random sample is at most $S_{(k)}$, and 0 otherwise (Bernoulli trials).

Let $X = \sum_{i=1}^{n^{3/4}} X_i$.

$$P[X_i = 1] = \frac{k}{n} \text{ and } E[X] = \frac{k}{n^{1/4}}$$

$$\sigma_X^2 = n^{3/4} \left(\frac{k}{n}\right) \left(1 - \frac{k}{n}\right) \leq \frac{n^{3/4}}{4} \text{ and } \sigma_X \leq \frac{n^{3/8}}{2}$$

Analysis of LazySelect II

Now we are ready to apply Chebyshev bounds on X.

$$P[a > S_{(k)}] = P[|X - E[X]| \geq \sqrt{n}] \leq P[|X - E[X]| \geq 2n^{1/8}\sigma_x] \leq \frac{1}{4n^{1/4}}.$$

A similar argument shows that $P[b < S_{(k)}] \leq \frac{1}{4n^{1/4}}$.

Case 2) We have to estimate the probability that P contains more than $4n^{3/4} + 2$ elements. This case can be done very similar to case 1) and is a nice question for your assignments.