Math 261C: Randomized Algorithms

Lecture topic: Floyd-Rivest Median Selection

Lecturer: Sam Buss
Scribe notes by: Marco Carmosino
Date: April 2, 2014

1. Improved Median Selection: The Algorithm

As sketched at the end of yesterday's lecture, we can improve our randomized median or (more generally) k^{th} element selection algorithm by skewing the pivot distribution towards elements that are close to the k^{th} element. The idea is to pick $\sqrt{n}$ elements at random and pivot using the k^{th} element of this subset, which should be close to the k^{th} element of A. This algorithm is orginally from [BFP$^+$73] and we follow the treatment from [Kiw05]. See [Pat96] for a survey of median finding.

Some notation and parameters:

- a_i^* is the i^{th} sorted element of A, and we
- s is the "sample size," the number of potential pivot points we sample from A
- g is the "gap" size

Data: A, n, k
Result: the k^{th} element of A
if *n = 1* **then**
  return a_0;
else
  Choose $S \subset A$ of size s uniformly at random without replacement ;
  $j_u = \max\{k \cdot \frac{s}{n} - g, 0\}$;
  $j_v = \min\{k \cdot \frac{s}{n} + g, 0\}$;
  ;
  $u = \text{FR-Select}(S, s, j_u)$;
  $v = \text{FR-Select}(S, s, j_v)$;
  ;
  `/* Scan` A `sequentially to partition it as below */`
  $\{u\}$;
  $\{v\}$;
  $L = \{a_i < u\}$;
  $M = \{u < a_i < v\}$;
  $U = \{a_i > v\}$;
  ;
  if $|L| = k$ **then**
    return u;
  else if $|L| + |M| + 1 = k$ **then**
    return v;
  else if $|L| > k$ **then**
    return FR-Select(L, $|L|$, k);
  else if $(|L| + |M| + 1) > k$ **then**
    return FR-Select(M, $|M|$, $k - |L| - 1$);
  else
    return FR-Select (U, $|U|$, $k - (|L| + |M| + 2)$);
  end
end

Algorithm 1: The Floyd-Rivest k^{th} element selection algorithm

2. The Analysis

As before, we will measure our runtime by expected number of comparison operations. The runtime will be dominated by scanning A sequentially, so the algorithm is at least linear. Further, without lose of generality, we can assume that $k < \frac{n}{2}$, because we can reverse A to arrange for this. We require some notation, let:

$$i_u \text{ be the index of } u \text{ such that } u = a^*_{i_u}$$
$$i_v \text{ be the index of } v \text{ such that } v = a^*_{i_v}$$

The algorithm compares first against u, and then against v if necessary.

We claim:

$$\begin{aligned}\mathbb{E}[\#\text{ comparisons}] &= n + i_v + (\#\text{ recursive calls})\\ &\leq n + \frac{n}{2} + o(n) + (\#\text{ recursive calls})\\ &= n + \frac{n}{2} + o(n)\end{aligned}$$

To begin, we will prove:

$$i_v \leq \frac{n}{2} + o(n)$$

The intuition here is that $i_v \approx j_v \cdot \frac{n}{s} \approx k + g \cdot \frac{n}{s}$, and $i_u \approx k - g \cdot \frac{n}{s}$. Specifically, we want to show that the following hold with high probability:

$$k - 2\frac{gn}{s} \underbrace{\leq}_{(1)} i_u \underbrace{\leq}_{(2)} k \underbrace{\leq}_{(3)} i_v \underbrace{\leq}_{(4)} k + 2\frac{gn}{s}$$

Note that:

- if (2) holds, we don't call FR-Select recursively on L
- if (2) and (3) hold, this implies that $|M| = o(n)$
- if (3) holds, we don't call FR-Select recursively on U
- if (4) holds, we have $i_v \leq \frac{n}{2} + o(1)$

Setting $s = \sqrt{n}$ and $g = n^{1/3}$, we have:

$$
\begin{aligned}
|M| &= i_v - i_u - 1 \\
&\leq 2 \cdot \frac{gn}{s} \\
&= 2n^{1/3} \cdot \frac{n}{n^{1/2}} \\
&= 2n^{5/6} \\
&= o(n)
\end{aligned}
$$

Lemma 1. Prob*[(3) fails] is o(1)*

Will will prove Lemma 1 above, but similar arguments show that (1), (2), (4) and (5) also fail with probability o(1).

Proof. Suppose $k > i_v$, then $v = j_v^{th}$ element of S and $v = a_{i_v}^*$, the i_v^{th} element of A in sorted order. Then $k > i_v$ iff the j_v^{th} element of S is greater than the k^{th} element of A, which occurs iff S has more than j_v many elements selected from the first k elements of A in sorted order.

We can describe the event above using a balls and urns model. The balls are members of A, and so there are n total balls. The red balls are $\{a_i : a_i < a_k^*\}$, and thus there are $\frac{k}{n}$ red balls total. To obtain the set S, we draw s balls form the urn. The *bad* event is that $> j_v$ of the balls drawn are red. Let's obtain an expression for j_v in terms of useful quantities:

$$
\begin{aligned}
j_v &= k \cdot \frac{s}{n} + g \\
&= (k + \frac{gn}{s})\frac{s}{n} \\
&= \frac{k}{n}
\end{aligned}
$$

Now, we use the following lemma, which improves Chernoff bounds for a balls-and-urn model:

Lemma 2 (Chvatal Chernoff Improvement [Chv79])**.** *If N balls have pN red with or without replacement, and M balls are drawn, then:*

$$\text{Prob}[> (p+t)M \textit{ balls in sample are red}] \leq e^{-2t^2 M}$$

We apply the bounds with: $M \leftarrow s$, $p \leftarrow \frac{k}{n}$, and $t \leftarrow \frac{g}{s}$, so:

$$
\begin{aligned}
\text{Prob}[(3) fails] &\le e^{-2t^2M} \\
&= e^{-2(g/s)^2 s} \\
&= e^{-2(g^2/s)} \\
&= o(1) && \text{by } s = n^{1/2} \text{ and } g = n^{1/2}
\end{aligned}
$$

□

We write out the runtime $T(n)$ in number of comparisons:

$$
\begin{aligned}
T(n) \le\; & n + \frac{n}{2} + o(1) \\
& + \text{Prob}[L \text{ or } U \text{ is recursed on}] \cdot T(\max\{|U|, |L|\}) \\
& + \text{Prob}[M \text{ is recursed on}] \cdot T(|M|) \\
& + T(\sqrt{n}) \\
\le\; & \frac{3}{2}n + o(n) + o(1) + T(n) + O(2n^{5/6}) \\
\le\; & \frac{3}{2}n + o(n)
\end{aligned}
$$

This completes our argument.

References

[BFP+73] Manuel Blum, Robert W. Floyd, Vaughan R. Pratt, Ronald L. Rivest, and Robert Endre Tarjan. Time bounds for selection. *J. Comput. Syst. Sci.*, 7(4):448–461, 1973.

[Chv79] Vašek Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.

[DZ99] Dorit Dor and Uri Zwick. Selecting the median. *SIAM J. Comput.*, 28(5):1722–1758, 1999.

[Kiw05] Krzysztof C. Kiwiel. On floyd and rivest's select algorithm. *Theor. Comput. Sci.*, 347(1-2):214–238, 2005.

[KL96] Rolf G. Karlsson and Andrzej Lingas, editors. *Algorithm Theory - SWAT '96, 5th Scandinavian Workshop on Algorithm Theory, Reykjavík, Iceland, July 3-5, 1996, Proceedings*, volume 1097 of *Lecture Notes in Computer Science*. Springer, 1996.

[Pat96] Mike Paterson. Progress in selection. In Karlsson and Lingas [KL96], pages 368–379.

[SPP76] Arnold Schönhage, Mike Paterson, and Nicholas Pippenger. Finding the median. *J. Comput. Syst. Sci.*, 13(2):184–199, 1976.