Asymptotic Analysis and Optimal Selection

Conrado Martínez
Univ. Polit. de Catalunya, Spain

Joint work with J. Daligault and R.M. Jiménez

CanaDAM, Montreal, May 2009

Introduction

Problem: Given an array A of n items and a rank m, $1 \leq m \leq n$, find the mth smallest element in A.
The algorithm should work in (expected) linear time $\Theta(n)$, irrespective of m.

Introduction

Hoare (1962) invents quickselect: pick some element p from the array, called the pivot, rearrange the contents of A so that all elements in A smaller that p are to its left, and all elements larger than p are to its right; if p is at position $j = m$ it is the sought element; if $j > m$ proceed recursively in $A[1..j-1]$, otherwise in $A[j+1..n]$.

The Quickselect Algorithm

```
Elem quickselect(vector<Elem>& A, int m) {
    int l = 0; int u = A.size() - 1;
    int k, p;
    while (l <= u) {
       p = select_pivot(A, l, u, m);
       swap(A[p], A[l]);
       partition(A, l, u, j);
       if (m < j) u = j-1;
       else if (m > j) l = j+1;
       else return A[j];
}   }
```

Analysis of Quickselect

Knuth (1971) shows that

$$\mathbb{E}[C_{n,m}] = 2\,\big(n+3+(n+1)H_n \\ -(m+2)H_m-(n+3-m)H_{n+1-m}\big)\,,$$

with $H_n = \sum_{1\le i\le n}(1/i) = \log n + \mathcal{O}(1)$ the nth harmonic number.

Analysis of Quickselect

- The expectation characteristic function:

$$f_1(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \frac{\mathbb{E}[C_{n,m}]}{n}$$

- The pth moment characteristic function:

$$f_p(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \frac{\mathbb{E}\left[C_{n,m}^p\right]}{n^p}$$

- For the variance we have

$$v(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \frac{\mathbb{V}[C_{n,m}]}{n^2} = f_2(\alpha) - f_1^2(\alpha)$$

Analysis of Quickselect

Example

- Standard quickselect:

$$f_1(\alpha) = 2 - 2 \cdot (\alpha \log \alpha + (1-\alpha)\log(1-\alpha)$$

- Median-of-three:

$$f_1(\alpha) = 2 + 3\alpha(1-\alpha)$$

Analysis of Quickselect

Example

- Standard quickselect:

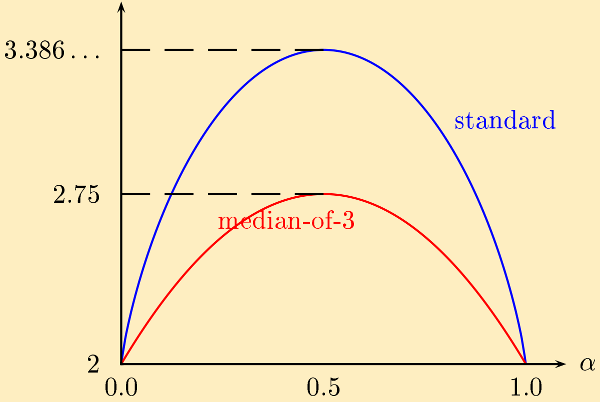
$$f_1(0) = f_1(1) = 2$$
$$f_1(1/2) = 2 + 2\log 2 \approx 3.386$$

- Median-of-three:

$$f_1(0) = f_1(1) = 2$$
$$f_1(1/2) = 11/4 = 2.75$$

Analysis of Quickselect

A plot of the standard quickselect characteristic function versus median-of-three characteristic function



Adaptive Sampling

- Adaptive sampling uses a sample of s elements to choose a pivot for each recursive stage of quickselect.
- If the current relative rank is $\alpha = m/n$, we select the element of rank $r(\alpha)$ from the sample

Example

- Standard quickselect: $s = 1, r(\alpha) = 1$
- Median-of-$(2t+1)$: $s = 2t + 1, r(\alpha) = t + 1$
- Proportion-from-s: $r(\alpha) \approx \alpha \cdot s$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

9	5	10	12	3	1	11	15	7	2	8	13	6	4	14

$$\alpha = 4/15 < 1/3$$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

9	5	10	12	3	1	11	15	7	2	8	13	6	4	14

$$\alpha = 4/15 < 1/3$$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

9	5	10	12	3	1	11	15	7	2	8	13	6	4	14

$$\alpha = 4/15 < 1/3$$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

7	5	4	6	3	1	8	2	9	15	11	13	12	10	14

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

7	5	4	6	3	1	8	2	9	15	11	13	12	10	14

$$1/3 < \alpha = 4/8 = 1/2 < 2/3$$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

7	5	4	6	3	1	8	2	9	15	11	13	12	10	14

$$1/3 < \alpha = 4/8 = 1/2 < 2/3$$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

1	5	4	2	3	6	8	7	9	15	11	13	12	10	14

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

1	5	4	2	3	6	8	7	9	15	11	13	12	10	14

$$\alpha = 4/5 > 2/3$$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

1	5	4	2	3	6	8	7	9	15	11	13	12	10	14

$$\alpha = 4/5 > 2/3$$

Adaptive Sampling

Example

We are looking the fourth element ($m = 4$) out of $n = 15$ elements

2	3	1	4	5	6	8	7	9	15	11	13	12	10	14

Adaptive Sampling

An adaptive sampling strategy can be characterized by the value of $r(\alpha)$ for a finite set of ℓ intervals that partition $[0,1]$, i.e., $r_k = r(\alpha)$ if $\alpha \in I_k$, $1 \le k \le \ell$.

The formal definition of adaptive sampling

$$
\begin{aligned}
& 0 = a_0 < a_1 < a_2 < \cdots < a_{\ell-1} < a_\ell = 1, \\
& I_1 = [0, a_1], \quad I_\ell = [a_{\ell-1}, 1], \\
& I_k = (a_{k-1}, a_k] \quad \text{if } k > 1 \text{ and } a_k \le 1/2, \\
& I_k = [a_{k-1}, a_k) \quad \text{if } k < \ell \text{ and } a_{k-1} > 1/2, \text{ and} \\
& I_k = (a_{k-1}, a_k) \quad \text{if } a_{k-1} \le 1/2 < a_k \text{ and } 1 < k < \ell.
\end{aligned}
$$

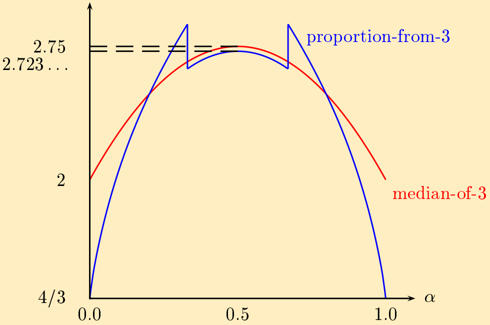
Adaptive Sampling

Example

- Standard quickselect: $s = 1; \ell = 1; r_1 = 1$
- Median-of-$(2t+1)$: $s = 2t+1; \ell = 1; r_1 = t+1$
- Proportion-from-s: $\ell = s; r_k = k$
- "Pure" proportion-from-s: proportion-from-s + $a_k = k/s$

Adaptive Sampling

A plot of median-of-three characteristic function versus proportion-from-three $f_1(\alpha)$



Adaptive Sampling

Theorem (Martínez, Panario, Viola (2004))

The expectation characteristic function $f(\alpha) \equiv f_1(\alpha)$ of any adaptive sampling strategy satisfies

$$
\begin{aligned}
f(\alpha) = 1 + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \\
&\left[\int_{\alpha}^{1} f(\alpha/x) x^{r(\alpha)} (1-x)^{s-r(\alpha)}\, dx \right. \\
&\quad \left. + \int_{0}^{\alpha} f\left(\frac{\alpha-x}{1-x}\right) x^{r(\alpha)-1} (1-x)^{s+1-r(\alpha)}\, dx \right].
\end{aligned}
$$

Adaptive Sampling

Theorem

The pth moment characteristic function $f_p(\alpha)$ of any adaptive sampling strategy satisfies

$$f_p(\alpha) = \psi_p(\alpha) + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \left[\int_\alpha^1 f_p(\alpha/x) x^{r(\alpha)+p-1} (1-x)^{s-r(\alpha)}\, dx + \int_0^\alpha f_p\left(\frac{\alpha - x}{1-x}\right) x^{r(\alpha)-1} (1-x)^{s+p-r(\alpha)}\, dx \right],$$

where

$$\psi_p(\alpha) = -(-1)^p \sum_{0 \le i < p} \binom{p}{i} (-1)^i f_i(\alpha), \qquad f_0(\alpha) = 1$$

Adaptive Sampling

Theorem

The pth moment characteristic function $f_p(\alpha)$ of any adaptive sampling strategy satisfies

$$f_p(\alpha) = \psi_p(\alpha) + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \left[\int_\alpha^1 f_p(\alpha/x) x^{r(\alpha)+p-1}(1-x)^{s-r(\alpha)}\,dx + \int_0^\alpha f_p\left(\frac{\alpha-x}{1-x}\right) x^{r(\alpha)-1}(1-x)^{s+p-r(\alpha)}\,dx\right],$$

where

$$\psi_p(\alpha) = -(-1)^p \sum_{0\le i<p} \binom{p}{i} (-1)^i f_i(\alpha), \qquad f_0(\alpha) = 1$$

Optimal Selection

Floyd and Rivest (1970) proposed an algorithm which uses sampling to obtain two pivots at each stage and achieves optimal expected performance.

$$\mathbb{E}[C_{n,m}] = n + \min(m, n - m) + \text{l.o.t.}$$

However, the algorithm is more complicated and uses samples of size $\Theta(n^{2/3} \log n)$; such a size seems to have been choosen for the proof to work

Quickselect with Large Samples

Theorem (Martínez, Panario, Viola (2004))

Biased proportion-from-s sampling with $s \to \infty$ achieves optimal expected performance:

$$f_1(\alpha) = 1 + \min(\alpha, 1 - \alpha)$$

Quickselect with Large Samples

- Intuition: Using very large sample and proportion-from-s helps, because we get a very good pivot, very close to the sought element; we can take $s = s(n)$ as long as $s(n) = o(n)$
- We should make sure that our pivot is very close BUT at the right side of the sought element! (i.e., slightly to the right if $\alpha < 1/2$, slightly to the left if $\alpha > 1/2$). That's what biased stands for in the previous theorem

Biased Sampling

Definition

A family of sampling strategies is biased if, for $\alpha < 1/2$,

$$r(\alpha) > s \cdot \alpha + 1 - \alpha$$

Quickselect with Large Samples

Theorem

For any biased proportion-from-s sampling with $s \to \infty$

$$f_p(\alpha) = (1 + \min(\alpha, 1 - \alpha))^p$$

In fact,

$$\frac{C_{n,m}}{n} \xrightarrow{d} 1 + \min(\alpha, 1 - \alpha),$$

as $n \to \infty$ *and* $m/n \to \alpha \in [0, 1]$

Quickselect with Large Samples

Theorem

For median-of-$(2t+1)$ sampling with $t \to \infty$

$$f_p(\alpha) = 2^p$$

Also,

$$\frac{C_{n,m}}{n} \xrightarrow{d} 2,$$

as $n \to \infty$ and $m/n \to \alpha \in [0,1]$

Sketch of the Proof

Some important facts to take into account for the analysis of quickselect with large samples

- We will consider only symmetric strategies:

$$r(\alpha) = s + 1 - r(1 - \alpha)$$

  For any symmetric strategy, $f_p(\alpha) = f_p(1 - \alpha)$
- The solution f_p of the integral equation is unique; the equation is of the form $f_p = T(f_p)$ and the operator T is a contraction
- For median-of-$(2t + 1)$, $f_p(0) = f_p(1) = 2^p + \mathcal{O}(1/t)$
- For proportion-from-s, if $r(\alpha) = 1$ for $\alpha \to 0$ then $f_p(0) = f_p(1) = 1 + p/s + \mathcal{O}(1/s^2)$

Sketch of the Proof

Our goal is to investigate the properties of the solution $f_p(\alpha)$ as $s \to \infty$,

$$\begin{aligned}
f_p(\alpha) = \psi_p(\alpha) + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \\
\times \left[\int_\alpha^1 f_p(\alpha/x) x^{r(\alpha)-1+p}(1-x)^{s-r(\alpha)}\, dx \right. \\
\left. + \int_0^\alpha f_p\left(\frac{\alpha-x}{1-x}\right) x^{r(\alpha)-1}(1-x)^{s-r(\alpha)+p}\, dx \right],
\end{aligned}$$

where

$$\psi_p(\alpha) = -(-1)^p \sum_{0 \le i < p} \binom{p}{i} (-1)^i f_i(\alpha), \qquad f_0(\alpha) = 1$$

Sketch of the Proof

In the right hand side, we have two integrals of the form

$$\int_a^b g(x)x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}dx$$

Recall:

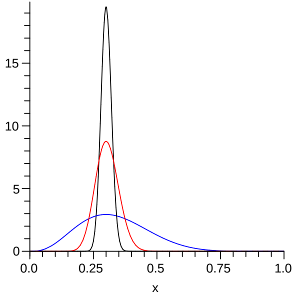
- For median-of-$(2t+1)$, $s = 2t+1$ and $r = t+1$
- For biased proportion-from-s, $r(\alpha) \approx \alpha s$

Sketch of the Proof

When $r, s \to \infty$

$$x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}$$

is highly concentrated around $x^* = (r-1)/(s-1)$



Sketch of the Proof

We can expect thus

$$\int_a^b g(x)x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}dx \to 0$$

if $x^* \notin [a, b]$, and

$$\int_a^b g(x)x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}dx$$
$$\sim \int_0^1 g(x)x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}dx,$$

if $x^* \in (a, b)$.
The case were $x^* = a$ or $x^* = b$ is slightly different.

Sketch of the Proof

Using Laplace's method we can show that

$$
\begin{aligned}
I(r,s) &= \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!}\int_a^b g(x)x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}dx \\
&= \begin{cases} g(x^*)+\mathcal{O}(1/s), & \text{if } a < x^* < b, \\ \mathcal{O}(1/s), & \text{if } x^* \notin [a,b] \\ g(x^*)/2+\mathcal{O}(1/s), & \text{if } x^* = a \text{ or } x^* = b \end{cases}
\end{aligned}
$$

provided g is in C^2 near x^*

Sketch of the Proof

We then show that $f_p(\alpha) = 2^p$ and $f_p(\alpha) = (1 + \min(\alpha, 1 - \alpha))^p$ are the (unique) fixed points for the integral equations corresponding to median-of-$(2t + 1)$ and biased proportion-from-s, respectively.

To do that we substitute our "guess" into the right hand side of the integral equation and use the asymptotic equivalents we've found before to show that indeed these are the solutions we sought.

The last part of the theorems follows after we check that the moments characterize the corresponding (deterministic!) limit distributions

Sketch of the Proof

For instance, for biased proportion-from-s if $\alpha < 2$, we have $x^* = (r-1)/(s-1) \to \alpha$, but $x^* > \alpha$,

$$\frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \left[\int_\alpha^1 f_p(\alpha/x) x^{r(\alpha)-1+p}(1-x)^{s-r(\alpha)}\,dx \right.$$
$$\left. + \int_0^\alpha f_p\left(\frac{\alpha - x}{1-x}\right) x^{r(\alpha)-1}(1-x)^{s-r(\alpha)+p}\,dx\right]$$
$$= f_p(\alpha/x^*)(x^*)^p + \mathcal{O}(1/s) = \alpha^p + \mathcal{O}(1/s)$$

Since $\psi_p(\alpha) = (1+\alpha)^p - \alpha^p$ for $\alpha < 1/2$, we prove $f_p(\alpha) = (1+\alpha)^p$ when $\alpha < 1/2$.

Final Remarks

The error terms in our asymptotic analysis allow us to prove that for proportion-from-s sampling with $s = s(n)$

$$\begin{aligned}\mathbb{E}[C_{n,m}] &= n + \min(m, n-m) + \mathcal{O}(s) + \mathcal{O}(n/s) \\ &\quad + \text{lower order terms independent of } s \\ \mathbb{V}[C_{n,m}] &= \Theta\left(\max\left(n \cdot s, \frac{n^2}{s}\right)\right)\end{aligned}$$

Therefore the optimal sample size is $s(n) = \Theta(\sqrt{n})$ as this simultaneously minimizes the lower order terms of the average cost and the order of magnitude of the variance

Thanks for your attention!