Adaptive Sampling for Quickselect

C. Martínez    D. Panario    A. Viola

Univ. Politècnica de Catalunya, Spain
Carleton University, Canada
Univ. de la República, Uruguay & LIPN - Université de Paris-Nord, France
Quickselect (Hoare, 1962) selects the $m$-th smallest element out of $n$ elements.
Quickselect (Hoare, 1962) selects the $m$-th smallest element out of $n$ elements. It partitions the given array around a pivot and continues into the appropriate subarray.
Introduction

Quickselect (Hoare, 1962) selects the $m$-th smallest element out of $n$ elements.

It partitions the given array around a pivot and continues into the appropriate subarray.

Quickselect is efficient: e.g. (Knuth, 1971)

$$C_{n,m} = m_0(\alpha) \cdot n + o(n) = 2(1+H(\alpha)) \cdot n + o(n)$$

$$= (2 - 2(\alpha \ln \alpha + (1 - \alpha) \ln(1 - \alpha))) \cdot n + o(n),$$

with $0 \leq \alpha = \frac{m}{n} \leq 1$. 
Elem quickselect(vector<Elem>& A, int m) {
    int l = 0; int u = A.size() - 1;
    int k, p;
    while (l <= u) {
        p = get_pivot(A, l, u, m);
        swap(A[p], A[l]);
        partition(A, l, u, k);
        if (m < k) u = k-1;
        else if (m > k) l = k+1;
        else return A[k];
    }
}
Median-of-$(2t + 1)$

Using a sample of $s = 2t + 1$ in each iteration improves the performance and reduces the probability of worst-case
Median-of-\((2t + 1)\)

Using a sample of \(s = 2t + 1\) in each iteration improves the performance and reduces the probability of worst-case.

For median-of-3 quickselect (Kirschenhofer, Martínez, Prodinger, 1995)

\[
m_1(\alpha) = 2 + 3\alpha(1 - \alpha).
\]
Median-of-\((2t + 1)\)

Using a sample of \(s = 2t + 1\) in each iteration improves the performance and reduces the probability of worst-case.

For median-of-3 quickselect (Kirschenhofer, Martínez, Prodinger, 1995)

\[
m_1(\alpha) = 2 + 3\alpha(1 - \alpha).
\]

For all \(\alpha, 0 \leq \alpha \leq 1\), \(m_0(\alpha) \leq m_1(\alpha)\). Also, \(\overline{m_0} = 3\) and \(\overline{m_1} = 2.5\).
Adaptive Sampling

Use the element in the sample with relative rank close to $\alpha = \frac{m}{n}$
Adaptive Sampling

Use the element in the sample with relative rank close to $\alpha = m/n$

In general: $r(\alpha) = \text{rank of the pivot within the sample, when selecting the } m\text{-th out of } n \text{ elements and } \alpha = m/n$
Adaptive Sampling

Use the element in the sample with relative rank close to $\alpha = m/n$

In general: $r(\alpha) = \text{rank of the pivot within the sample, when selecting the } m\text{-th out of } n \text{ elements and } \alpha = m/n$

Divide $[0, 1]$ into $\ell$ intervals with endpoints

$$0 = a_0 < a_1 < a_2 < \cdots < a_\ell = 1$$

and let $r_k$ denote the value of $r(\alpha)$ for $\alpha$ in the $k$-th interval
Adaptive Sampling

For median-of-$(2t + 1)$: $\ell = 1$ and $r_1 = t + 1$
Adaptive Sampling

For median-of-$(2t + 1)$: $\ell = 1$ and $r_1 = t + 1$

For proportion-from-$s$: $\ell = s$, $a_k = k/s$ and $r_k = k$
Adaptive Sampling

- For median-of-$(2t + 1)$: $\ell = 1$ and $r_1 = t + 1$
- For proportion-from-$s$: $\ell = s$, $a_k = k/s$ and $r_k = k$
- “Proportion-from”-like strategies: $\ell = s$ and $r_k = k$, but the endpoints of the intervals $a_k \neq k/s$
Adaptive Sampling

- For median-of-\((2t + 1)\): \(\ell = 1\) and \(r_1 = t + 1\)
- For proportion-from-\(s\): \(\ell = s\), \(a_k = k/s\) and \(r_k = k\)
- “Proportion-from”-like strategies: \(\ell = s\) and \(r_k = k\), but the endpoints of the intervals \(a_k \neq k/s\)
- A sampling strategy is **symmetric** if
  \[
  r(\alpha) = s + 1 - r(1 - \alpha).
  \]
The Recurrence

Probability that the $r$-th element in a sample of size $s$ is the $j$-th element of the $n$ given elements:

$$\pi_{n,j}^{(s,r)} = \frac{(j-1) \binom{n-j}{s-r}}{\binom{n}{s}}$$

$$1 \leq r \leq s \leq n, \quad 1 \leq j \leq n.$$
The Recurrence

Average number of comparisons $C_{n,m}$ to select the $m$-th out of $n$:

$$C_{n,m} = n + \Theta(1) + \sum_{j=m+1}^{n} \pi_{n,j}^{(s,r)} \cdot C_{j-1,m}$$

$$+ \sum_{j=1}^{m-1} \pi_{n,j}^{(s,r)} \cdot C_{n-j,m-j}.$$
A General Theorem

Theorem 1. Let $f(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n}$. Then

$$f(\alpha) = 1 + \frac{s!}{(r(\alpha) - 1)!(s - r(\alpha))!} \times \left[ \int_{\alpha}^{1} f \left( \frac{\alpha}{x} \right) x^{r(\alpha)} (1 - x)^{s-r(\alpha)} \, dx \right. \right. \left. \left. + \int_{0}^{\alpha} f \left( \frac{\alpha - x}{1 - x} \right) x^{r(\alpha)-1} (1 - x)^{s+1-r(\alpha)} \, dx \right] . $$
Two Elementary Facts

If $r(\alpha)$ is symmetric then $f(\alpha) = f(1 - \alpha)$. 

Let $r_0 = \lim_{\tau \to 0} r(\tau)$. Then $\lim_{\tau \to 0} f(\tau) = s + 1$, with $r_0 = \frac{1}{s + 1}$; hence, $f(0) = 1 + 1 = s$, while for median-of-$(2t + 1)$, we have $m_t(0) = 2$. 

In proportion-from strategies $r_0 = 1$; hence, $f(0) = 1 + 1 = s$, while for median-of-$(2t + 1)$, we have $m_t(0) = 2$. 

ALCOM-FT Final Review, Árhus, Denmark, January 2004
Two Elementary Facts

- If $r(\alpha)$ is symmetric then $f(\alpha) = f(1 - \alpha)$.
- Let $r_0 = \lim_{\alpha \to 0} r(\alpha)$. Then
  \[
  \lim_{\alpha \to 0} f(\alpha) = \frac{s + 1}{s + 1 - r_0}.
  \]

  In proportion-from strategies $r_0 = 1$; hence, $f(0) = 1 + 1/s$, while for median-of-$(2t + 1)$, we have $m_t(0) = 2$.
Denote \( f_k \) the restriction of \( f(\alpha) \) to the \( k \)-th interval of \([0, 1]\).

**Lemma 1.** For any adaptive sampling strategy

\[
\frac{d^{s+2}}{d\alpha^{s+2}} f_k(\alpha) = \frac{(-1)^{s+1-r_k} \cdot s!}{\alpha^{s+1-r_k} (r_k - 1)!} \frac{d^{r_k+1}}{d\alpha^{r_k+1}} f_k(\alpha) \\
+ \frac{s!}{(1 - \alpha)^{r_k} (s - r_k)!} \frac{d^{s+2-r_k}}{d\alpha^{s+2-r_k}} f_k(\alpha).
\]
Two Problems and a Trick

Solving high-order linear differential equations
Two Problems and a Trick

- Solving high-order linear differential equations
- We do not know the initial values of the $f_k$’s and their derivatives
Solving high-order linear differential equations

We do not know the initial values of the $f_k$'s and their derivatives

Plug the general form of the $f_k$’s back into the integral equation(s) and solve for the unknown constants
The differential equation is

\[ \frac{d^2 \phi_1}{dx^2} - \frac{2}{1 - x} \frac{d\phi_1}{dx} - \frac{2}{x^2} \phi_1 = 0 \]

with \( \phi_1(x) = f_1''(x) \) and \( f_2(x) = f_1(1 - x) \).
The differential equation is

\[ \frac{d^2 \phi_1}{dx^2} - \frac{2}{1 - x} \frac{d \phi_1}{dx} - \frac{2}{x^2} \phi_1 = 0 \]

with \( \phi_1(x) = f''_1(x) \) and \( f_2(x) = f_1(1 - x) \).

The solution is

\[ f_1(x) = a \left( (x - 1) \ln(1 - x) + \frac{x^3}{6} + \frac{x^2}{2} - x \right) - b(1 + \mathcal{H}(x)) + cx + d. \]
The maximum is at $\alpha = 1/2$. There

$$f(1/2) = 3.112 \ldots$$
The maximum is at $\alpha = 1/2$. There
\[ f(1/2) = 3.112\ldots \]
Proportion-from-2 beats standard quickselect:
\[ f(\alpha) \leq m_0(\alpha) \]
The maximum is at $\alpha = 1/2$. There
$f(1/2) = 3.112 \ldots$

Proportion-from-2 beats standard quickselect:
$f(\alpha) \leq m_0(\alpha)$

Proportion-from-2 beats median-of-three in
some regions: $f(\alpha) \leq m_1(\alpha)$ if $\alpha \leq 0.140 \ldots$ or
$\alpha \geq 0.860 \ldots$
The maximum is at $\alpha = 1/2$. There
\[ f(1/2) = 3.112 \ldots \]
Proportion-from-2 beats standard quickselect:
\[ f(\alpha) \leq m_0(\alpha) \]
Proportion-from-2 beats median-of-three in some regions:
\[ f(\alpha) \leq m_1(\alpha) \text{ if } \alpha \leq 0.140 \ldots \text{ or } \alpha \geq 0.860 \ldots \]
The grand-average: \[ C_n = \overline{f} \cdot n + o(n), \text{ with } \overline{f} = 2.598 \ldots \]
Proportion-from-2

$m_0(\alpha)$

$m_1(\alpha)$

$f(\alpha)$

\[
\begin{align*}
&3.386 \ldots \\
&3.113 \ldots \\
&2.75 \quad 2 \\
&1.5 \\
&0.0 & 0.5 & 1.0
\end{align*}
\]
Proportion-from-3

For proportion-from-3,

\[ f_1(x) = -C_0(1 + H(x)) + C_1 + C_2x + C_3K_1(x) + C_4K_2(x), \]
\[ f_2(x) = -C_5(1 + H(x)) + C_6x(1 - x) + C_7, \]

with

\[ K_1(x) = \cos(\sqrt{2} \ln x) \cdot \sum_{n \geq 0} A_n x^{n+4} + \sin(\sqrt{2} \ln x) \cdot \sum_{n \geq 0} B_n x^{n+4}, \]
\[ K_2(x) = \sin(\sqrt{2} \ln x) \cdot \sum_{n \geq 0} A_n x^{n+4} - \cos(\sqrt{2} \ln x) \cdot \sum_{n \geq 0} B_n x^{n+4}. \]
Two maxima at \( \alpha = 1/3 \) and \( \alpha = 2/3 \). There
\[
f(1/3) = f(2/3) = 2.883 \ldots
\]
Two maxima at $\alpha = 1/3$ and $\alpha = 2/3$. There
\[ f(1/3) = f(2/3) = 2.883 \ldots \]

The median is not the most difficult rank:
\[ f(1/2) = 2.723 \ldots \]
Two maxima at $\alpha = 1/3$ and $\alpha = 2/3$. There $f(1/3) = f(2/3) = 2.883 \ldots$

The median is not the most difficult rank: $f(1/2) = 2.723 \ldots$

Proportion-from-3 beats median-of-three in some regions: $f(\alpha) \leq m_1(\alpha)$ if $\alpha \leq 0.201 \ldots$, $\alpha \geq 0.798 \ldots$ or $1/3 < \alpha < 2/3$
Proportion-from-3

Two maxima at $\alpha = 1/3$ and $\alpha = 2/3$. There

$$f(1/3) = f(2/3) = 2.883\ldots$$

The median is not the most difficult rank:

$$f(1/2) = 2.723\ldots$$

Proportion-from-3 beats median-of-three in some regions: $f(\alpha) \leq m_1(\alpha)$ if $\alpha \leq 0.201\ldots$, $\alpha \geq 0.798\ldots$ or $1/3 < \alpha < 2/3$

The grand-average: $C_n = \overline{f} \cdot n + o(n)$, with

$$\overline{f} = 2.421\ldots$$
Proportion-from-3: Batfind

\[ f(\alpha) \]

\[ m_1(\alpha) \]

\[ \alpha \]

\[ \frac{4}{3} \]

2.75

2.723\ldots

0.0 0.202\ldots 0.5 1.0

0.276\ldots
Proportion-from-3: Batfind

\[ f(\alpha) \]

![Graph showing the function \( f(\alpha) \)]
Like proportion-from-3, but $a_1 = \nu$ and $a_2 = 1 - \nu$
ν-find

Like proportion-from-3, but $a_1 = \nu$ and $a_2 = 1 - \nu$

Same differential equation, same $f_i$’s, with $C_i = C_i(\nu)$
Like proportion-from-3, but $a_1 = \nu$ and

$a_2 = 1 - \nu$

Same differential equation, same $f_i$’s, with

$C_i = C_i(\nu)$

If $\nu \rightarrow 0$ then $f_\nu \rightarrow m_1$ (median-of-three)
Like proportion-from-3, but $a_1 = \nu$ and $a_2 = 1 - \nu$

Same differential equation, same $f_i$’s, with $C_i = C_i(\nu)$

If $\nu \to 0$ then $f_{\nu} \to m_1$ (median-of-three)

However, if $\nu \to 1/2$ then $f_{\nu}$ behaves like proportion-from-2, but it is not the same
The optimal $\nu$

**Theorem 2.** There exists a value $\nu^*$, namely, $\nu^* = 0.182 \ldots$, such that for any $\nu$, $0 < \nu < 1/2$, and any $\alpha$,

$$f_{\nu^*}(\alpha) \leq f_{\nu}(\alpha).$$

Furthermore, $\nu^*$ is the unique value of $\nu$ such that $f_{\nu}$ is continuous, i.e.,

$$f_{\nu^*,1}(\nu^*) = f_{\nu^*,2}(\nu^*).$$
If \( \nu > \tilde{\nu} = 0.268 \ldots \) then \( f_\nu \) has two absolute maxima at \( \alpha = \nu \) and \( \alpha = 1 - \nu \); otherwise there is one absolute maximum at \( \alpha = 1/2 \).
More on $\nu$-find

If $\nu > \tilde{\nu} = 0.268 \ldots$ then $f_\nu$ has two absolute maxima at $\alpha = \nu$ and $\alpha = 1 - \nu$; otherwise there is one absolute maximum at $\alpha = 1/2$.

Obviously, the value $\nu^*$ minimizes the maximum

$$f_{\nu^*}(1/2) = 2.659 \ldots$$

and the mean

$$\bar{f}_{\nu^*} = 2.342 \ldots$$
More on $\nu$-find
More on $\nu$-find

If $\nu \leq \bar{\nu}' = 0.404 \ldots$ then $\nu$-find beats median-of-3 on average ranks: $\bar{f}_\nu \leq 5/2$
More on $\nu$-find

- If $\nu \leq \overline{\nu}' = 0.404 \ldots$ then $\nu$-find beats median-of-3 on average ranks: $\overline{f}_\nu \leq 5/2$

- If $\nu \leq \nu'_m = 0.364 \ldots$ then $\nu$-find beats median-of-3 to find the median: $f_\nu(1/2) \leq 11/4$
More on \( \nu \)-find

- If \( \nu \leq \nu' = 0.404 \ldots \) then \( \nu \)-find beats median-of-3 on average ranks: \( \bar{f}_\nu \leq 5/2 \)

- If \( \nu \leq \nu'_m = 0.364 \ldots \) then \( \nu \)-find beats median-of-3 to find the median:
  \[ f_\nu(1/2) \leq 11/4 \]

- If \( \nu \leq \nu' = 0.219 \ldots \) then \( \nu \)-find beats median-of-3 for all ranks:
  \[ f_\nu(\alpha) \leq m_1(\alpha) \]
More on $\nu$-find

$\mathbf{f}_{1,\nu}(\nu)$

$\mathbf{f}_{2,\nu}(\nu)$
More on $\nu$-find

$\nu^\ast$ and $\tilde{\nu}$

$f_{1,\nu}(\nu)$ and $f_{2,\nu}(\nu)$
More on $\nu$-find

$\mathcal{M}_0(\nu)$

$f_{1,\nu}(\nu)$

$f_{2,\nu}(\nu)$
More on $\nu$-find
More on $\nu$-find

\[ f_{1,\nu}(\nu) \]
\[ f_{2,\nu}(\nu) \]
\[ m_1(\nu) \]
More on $\nu$-find

The diagram shows a graph with axis labels $\nu$ and $f_\nu(1/2)$, with a range of $\nu$ from 0.0 to 0.5 and values of $f_\nu(1/2)$ from 2.66 to 3.1. The $m_1(1/2)$ line intersects the graph at specific points along the $\nu$ axis.
More on $\nu$-find

We have investigated the average total cost of $\nu$-find

$$\lambda_1 \cdot \# \text{ of comparisons} + \lambda_2 \cdot \# \text{ of exchanges}$$

The values of $\nu^*$ (optimum), $\nu'$ ($\nu$-find beats median-of-three), etc. now depend on $\lambda_2/\lambda_1$; for instance, if $\lambda_2/\lambda_1 = \infty$ we minimize the average number of exchanges with $\nu^* = 0.43 \ldots$
Proportion-from-s: Sharkfind
Theorem 3. Let $f_s(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n}$ when using samples of size $s$. Then for any adaptive sampling strategy such that $\lim_{s \to \infty} r(\alpha)/s = \alpha$

$$f_\infty(\alpha) = \lim_{s \to \infty} f_s(\alpha) = 1 + \min(\alpha, 1 - \alpha).$$

This is theoretically optimal for comparison-based selection algorithms.