Adaptive Sampling for Selection

Conrado Martínez

Joint work with:
J. Daligault (ENS Cachan, France)
D. Panario (Carleton U., Canada)
A. Viola (U. República, Uruguay)

Univ. Politècnica de Catalunya, Spain

February, 2006
1 Introduction

2 Small Samples

3 Large Samples
Problem: Given an array $A$ of $n$ items and a rank $m$, $1 \leq m \leq n$, find the $m$th smallest element in $A$. The algorithm should work in (expected) linear time $\Theta(n)$, irrespective of $m$. 
Hoare (1962) invents \textit{quickselect}: pick some element $p$ from the array, called the \textit{pivot}, rearrange the contents of $A$ so that all elements in $A$ smaller than $p$ 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 $A[j + 1..n]$. 
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];
    }
}
The expectation characteristic function:

\[ f(\alpha) = \lim_{\frac{m}{n} \to \alpha} \frac{\mathbb{E}[C_{n,m}]}{n} \]

The second factorial moment characteristic function:

\[ g(\alpha) = \lim_{\frac{m}{n} \to \alpha} \frac{\mathbb{E}[C_{n,m}^2]}{n^2} \]

For the variance we have

\[ \nu(\alpha) = \lim_{\frac{m}{n} \to \alpha} \frac{\mathbb{V}[C_{n,m}]}{n^2} = g(\alpha) - f^2(\alpha) \]
The expectation characteristic function:

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

The second factorial moment characteristic function:

\[ g(\alpha) = \lim_{\substack{n \to \infty \\mbox{m/n \to \alpha}}} \frac{\mathbb{E}[C_{n,m}^2]}{n^2} \]

For the variance we have

\[ v(\alpha) = \lim_{\substack{n \to \infty \\mbox{m/n \to \alpha}}} \frac{\mathbb{V}[C_{n,m}]}{n^2} = g(\alpha) - f^2(\alpha) \]
• The expectation characteristic function:

\[ f(\alpha) = \lim_{n \to \infty} \frac{\mathbb{E}[C_{n,m}]}{n} \]

• The second factorial moment characteristic function:

\[ g(\alpha) = \lim_{n \to \infty} \frac{\mathbb{E}[C_{n,m}^2]}{n^2} \]

• For the variance we have

\[ \nu(\alpha) = \lim_{n \to \infty} \frac{\mathbb{V}[C_{n,m}]}{n^2} = g(\alpha) - f^2(\alpha) \]
Example

- Standard quickselect (Knuth, 1971):
  \[ f(\alpha) = m_0(\alpha) = 2 - 2(\alpha \ln \alpha + (1 - \alpha) \ln(1 - \alpha)) = 2 + 2 \cdot H(\alpha) \]

- Median-of-three (Kirschenhofer, Martínez & Prodinger, 1997):
  \[ f(\alpha) = m_1(\alpha) = 2 + 3\alpha(1 - \alpha) \]
A plot of the standard quickselect and median-of-three characteristic functions.
Introduction

Small Samples

Large Samples
• 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 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-$s$: $r(\alpha) \approx \alpha \cdot s$
Example

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

\[
\begin{array}{cccccccccccccc}
9 & 5 & 10 & 12 & 3 & 1 & 11 & 15 & 7 & 2 & 8 & 13 & 6 & 4 & 14 \\
\end{array}
\]

\(\alpha = \frac{4}{15} < \frac{1}{3}\)
Example

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

\[
\begin{array}{cccccccccccccc}
9 & 5 & 10 & 12 & 3 & 1 & 11 & 15 & 7 & 2 & 8 & 13 & 6 & 4 & 14 \\
\end{array}
\]

\[
\alpha = \frac{4}{15} < \frac{1}{3}
\]
Example

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

\[
\begin{array}{cccccccccccccc}
9 & 5 & 10 & 12 & 3 & 1 & 11 & 15 & 7 & 2 & 8 & 13 & 6 & 4 & 14 \\
\end{array}
\]

\[\alpha = \frac{4}{15} < \frac{1}{3}\]
Example

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

\[
\begin{array}{cccccccccccccc}
7 & 5 & 4 & 6 & 3 & 1 & 8 & 2 & 9 & 15 & 11 & 13 & 12 & 10 & 14
\end{array}
\]
Example

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

<table>
<thead>
<tr>
<th>7</th>
<th>5</th>
<th>4</th>
<th>6</th>
<th>3</th>
<th>1</th>
<th>8</th>
<th>2</th>
<th>9</th>
<th>15</th>
<th>11</th>
<th>13</th>
<th>12</th>
<th>10</th>
<th>14</th>
</tr>
</thead>
</table>

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

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

\[
\begin{array}{cccccccccc}
7 & 5 & 4 & 6 & 3 & 1 & 8 & 2 & 9 & 15 & 11 & 13 & 12 & 10 & 14 \\
\end{array}
\]

\[
1/3 < \alpha = 4/8 = 1/2 < 2/3
\]

Adaptive Sampling for Selection

Data Structures 2006
Example

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

\[
\begin{array}{cccccccccccccc}
1 & 5 & 4 & 2 & 3 & 6 & 8 & 7 & 9 & 15 & 11 & 13 & 12 & 10 & 14
\end{array}
\]
Example

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

\[
\begin{array}{cccccccccccccc}
1 & 5 & 4 & 2 & 3 & 6 & 8 & 7 & 9 & 15 & 11 & 13 & 12 & 10 & 14 \\
\end{array}
\]

\(\alpha = 4/5 > 2/3\)
Example

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

\[
\begin{array}{cccccccccccccc}
1 & 5 & 4 & 2 & 3 & 6 & 8 & 7 & 9 & 15 & 11 & 13 & 12 & 10 & 14 \\
\end{array}
\]

\[\alpha = \frac{4}{5} > \frac{2}{3}\]
Example

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

\[
\begin{array}{cccccccccccccc}
2 & 3 & 1 & 4 & 5 & 6 & 8 & 7 & 9 & 15 & 11 & 13 & 12 & 10 & 14 \\
\end{array}
\]
A plot of standard, median-of-three and proportion-from-two characteristic functions.
A plot of median-of-three versus Batfind (a.k.a. proportion-from-three) characteristic functions
- \( \nu\text{-find} \) is like proportion-from-3, but cut points located at \( \nu \) and \( 1 - \nu \), instead of \( 1/3 \) and \( 2/3 \)
- If \( \nu \to 0 \) then \( f_\nu \to m_1 \) (median-of-three)
- If \( \nu \to 1/2 \) then \( f_\nu \) behaves like proportion-from-2, but it is not the same
• \( \nu \)-find is like proportion-from-3, but cut points located at \( \nu \) and \( 1 - \nu \), instead of \( 1/3 \) and \( 2/3 \)

• If \( \nu \to 0 \) then \( f_\nu \to m_1 \) (median-of-three)

• If \( \nu \to 1/2 \) then \( f_\nu \) behaves like proportion-from-2, but it is not the same
- \( \nu\text{-find} \) is like proportion-from-3, but cut points located at \( \nu \) and \( 1 - \nu \), instead of \( 1/3 \) and \( 2/3 \)

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

- If \( \nu \to 1/2 \) then \( f_\nu \) behaves like proportion-from-2, but it is not the same
A plot of \( \nu \)-find's characteristic function for various values of \( \nu \)

\[
f_{1/3}(\alpha)
\]

\( \alpha \)

\( 0.1 \quad 0.3 \quad 0.5 \quad 0.7 \)
A plot of $\nu$-find's characteristic function for various values of $\nu$.
A plot of \( \nu\text{-find}'s \) characteristic function for various values of \( \nu \)

\[
f_{1/5}(\alpha)
\]
A plot of $\nu$-find's characteristic function for various values of $\nu$ 

$$f_{\nu^*}(\alpha)$$
A plot of $\nu$-find's characteristic function for various values of $\nu$

$$f_{1/7}(\alpha)$$

$\alpha$

0.1 0.3 0.5 0.7
A plot of $\nu$-find's characteristic function for various values of $\nu$
A 3D plot of $nu$-find's characteristic function.
Theorem

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.,

$$f_{\nu^*,1}(\nu^*) = f_{\nu^*,2}(\nu^*)$$
Theorem

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 we consider average total cost then \( \nu^* \approx 0.25 \).
Theorem

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

$$f(\alpha) = 1 + \frac{s!}{(r(\alpha) - 1)!(s - r(\alpha))!} \times$$

$$\left[ \int_{\alpha}^{1} f \left( \frac{\alpha}{x} \right) x^{r(\alpha)}(1 - x)^{s-r(\alpha)} \, dx \right.$$  

$$+ \int_{0}^{\alpha} f \left( \frac{\alpha - x}{1 - x} \right) x^{r(\alpha)-1}(1 - x)^{s+1-r(\alpha)} \, dx \right]$$
Theorem

The second factorial moment characteristic function $g(\alpha)$ of any adaptive sampling strategy satisfies

$$g(\alpha) = 2f(\alpha) - 1 + \frac{s!}{(r(\alpha) - 1)!(s - r(\alpha))!} \times$$

$$\left[ \int_{\alpha}^{1} g\left(\frac{\alpha}{x}\right) x^{r(\alpha) + 1} (1 - x)^{s - r(\alpha)} \, dx + \int_{0}^{\alpha} g\left(\frac{\alpha - x}{1 - x}\right) x^{r(\alpha) - 1} (1 - x)^{s + 2 - r(\alpha)} \, dx \right]$$
A plot of $v(\alpha)$ for standard quickselect (Kirschenhofer, Prodinger (1998)) and for median-of-three.
1. Introduction

2. Small Samples

3. 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 should make sure that our pivot is very close to the right side of the sought element! (i.e., slightly to the right if $\alpha < 1/2$, slightly to the left if $\alpha > 1/2$)
• Intuition: Using very large sample and proportion-from-s helps, because we get a very good pivot, very close to the sought element

• 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 $\alpha > 1/2$)
**Definition**

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

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

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

\[ f(\alpha) = 1 + \min(\alpha, 1 - \alpha) \]
The proof of Martínez, Panario, Viola (2004) for fixed-size sampling with \( s \to \infty \) works also for variable-size samples, i.e., \( s = s(n) \), as long as \( s(n) \to \infty \) and \( s(n)/n \to 0 \) when \( n \to \infty \).

**Theorem**

For biased proportion-from-\( s \) sampling with increasing variable-size samples (i.e., \( s = s(n) \to \infty, s/n \to 0 \)),

\[
\mathbb{E}[C_{n,m}] = n + \min(m, n - m) + \Theta \left( \max \left( s, \frac{n}{s} \right) \right)
\]
Theorem

Biased proportion-from-$s$ sampling with $s \to \infty$ has subquadratic variance:

$$v(\alpha) = \lim_{n \to \infty} \frac{\text{Var}[C_{n,m}]}{n^2} = 0$$

$$m/n \to \alpha$$
Theorem

Biased proportion-from-$s$ sampling with $s \to \infty$ has subquadratic variance:

$$v(\alpha) = \lim_{{n \to \infty}} \lim_{{m/n \to \alpha}} \frac{\text{Var}[C_{n,m}]}{n^2} = 0$$

The same holds true for median-of-$(2t+1)$, when $t \to \infty$
Theorem

For biased proportion-from-s sampling with increasing variable-size samples (i.e., \( s = s(n) \to \infty, s/n \to 0 \)),

\[
\forall[C_{n,m}] = \Theta \left( \max \left( \frac{n^2}{s}, n \cdot s \right) \right)
\]
Theorem

The optimal sample size to minimize both the variance and the expected value of proportion-from-s is

\[ s^* = \Theta(\sqrt{n}) \]
Open Problems

- Obtain explicit solutions for interesting particular cases
- Show that for any fixed sample size $s$, the optimal strategy is proportion-from-$s$ at appropriate cut points (like $v$-find)
- Find a simple (exact or approximate) formula for the location of optimal cut points
- Why the coefficient of $n^2$ in the variance of median-of-3 is bimodal? Any intuitive explanation?
- Better asymptotic estimates for the optimal sample size $s^*$ (namely, the coefficient of $\sqrt{n}$). How does it depend on $\alpha$?
Sources
