On the Variance of Quickselect

Jean Daligault[1] Conrado Martínez[2]

[1] ENS Cachan, France

[2] Univ. Politècnica de Catalunya, Spain

January, 2006

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.

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 $A[j+1..n]$.

```
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];
}   }
```

Knuth (1971) shows that

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

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

- The expectation characteristic function:

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

- The second factorial moment characteristic function:

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

- For the variance we have

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

- The expectation characteristic function:

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

- The second factorial moment characteristic function:

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

- For the variance we have

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

- The expectation characteristic function:

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

- The second factorial moment characteristic function:

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

- For the variance we have

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

Example

- Standard quickselect:

$$f(\alpha) = m_0(\alpha) = 2-2(\alpha \ln \alpha + (1-\alpha)\ln(1-\alpha)) = 2+2\cdot\mathcal{H}(\alpha)$$

- Median-of-three:

$$f(\alpha) = m_1(\alpha) = 2 + 3\alpha(1-\alpha)$$

Example

- Standard quickselect:

$$f(\alpha) = m_0(\alpha) = 2-2(\alpha \ln \alpha + (1-\alpha)\ln(1-\alpha)) = 2+2\cdot\mathcal{H}(\alpha)$$

- Median-of-three:

$$f(\alpha) = m_1(\alpha) = 2 + 3\alpha(1-\alpha)$$

Example

- Standard quickselect:

$$m_0(0) = m_0(1) = 2$$
$$m_0(1/2) = 2 + 2\ln 2 \approx 3.386$$

- Median-of-three:

$$m_1(0) = m_1(1) = 2$$
$$m_1(1/2) = 11/4 = 2.75$$

Example

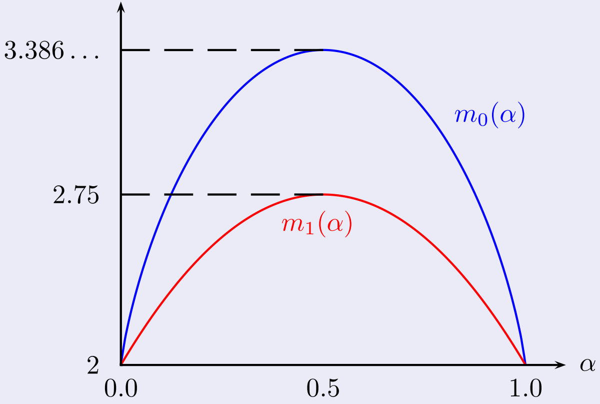
- Standard quickselect:

$$m_0(0) = m_0(1) = 2$$
$$m_0(1/2) = 2 + 2\ln 2 \approx 3.386$$

- Median-of-three:

$$m_1(0) = m_1(1) = 2$$
$$m_1(1/2) = 11/4 = 2.75$$

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



- 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: $r(\alpha) \approx \alpha \cdot s$

- 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\alpha \in I_k$, $1 \leq k \leq \ell$.

$$0 = a_0 < a_1 < a_2 < \cdots < a_{\ell-1} < a_\ell = 1,$$

$$I_1 = [0, a_1], \quad I_\ell = [a_{\ell-1}, 1],$$

$$I_k = (a_{k-1}, a_k] \quad \text{if } k > 1 \text{ and } a_k \leq 1/2,$$

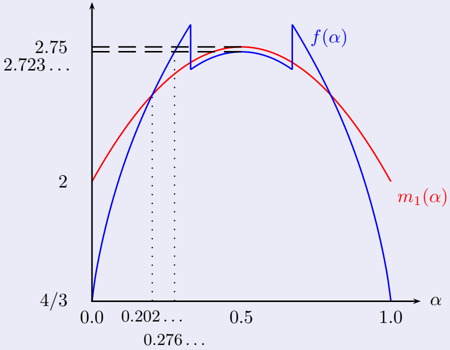
$$I_k = [a_{k-1}, a_k) \quad \text{if } k < \ell \text{ and } a_{k-1} > 1/2, \text{ and}$$

$$I_k = (a_{k-1}, a_k) \quad \text{if } a_{k-1} \leq 1/2 < a_k \text{ and } 1 < k < \ell.$$

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 + $a_k = k/s$

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



Theorem (Martínez, Panario, Viola (2004))

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

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

Lemma (Martínez, Panario, Viola (2004))

Let f_k be the restriction of $f(\alpha)$ to the kth interval I_k, and $r_k = r(\alpha)$ when $\alpha \in I_k$. For any adaptive sampling strategy

$$\frac{d^{s+2}}{d\alpha^{s+2}} f_k(\alpha) = \frac{(-1)^{s+1-r_k} \cdot s!}{\alpha^{s+1-r_k}(r_k - 1)!} \frac{d^{r_k+1}}{d\alpha^{r_k+1}} f_k(\alpha) + \frac{s!}{(1-\alpha)^{r_k}(s-r_k)!} \frac{d^{s+2-r_k}}{d\alpha^{s+2-r_k}} f_k(\alpha).$$

Theorem (Martínez, Panario, Viola (2004))

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

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

Theorem

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

$$g(\alpha) = 2f(\alpha) - 1 + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!}\left[\int_\alpha^1 g(\alpha/x)x^{r(\alpha)+1}(1-x)^{s-r(\alpha)}\,dx + \int_0^\alpha g\left(\frac{\alpha-x}{1-x}\right)x^{r(\alpha)-1}(1-x)^{s+2-r(\alpha)}\,dx\right].$$

Theorem

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

$$g(\alpha) = 2f(\alpha) - 1 + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!}\left[\int_\alpha^1 g(\alpha/x)x^{r(\alpha)+1}(1-x)^{s-r(\alpha)}\,dx + \int_0^\alpha g\left(\frac{\alpha-x}{1-x}\right)x^{r(\alpha)-1}(1-x)^{s+2-r(\alpha)}\,dx\right].$$

Lemma

Let g_k be the restriction of $g(\alpha)$ to the kth interval I_k, and $r_k = r(\alpha)$ when $\alpha \in I_k$. For any adaptive sampling strategy

$$\frac{d^{s+3}}{d\alpha^{s+3}} g_k(\alpha) = 2\frac{d^{s+3}}{d\alpha^{s+3}} f_k(\alpha) + \frac{(-1)^{s+1-r_k} \cdot s!}{\alpha^{s+1-r_k}(r_k-1)!} \frac{d^{r_k+2}}{d\alpha^{r_k+2}} g_k(\alpha) + \frac{s!}{(1-\alpha)^{r_k}(s-r_k)!} \frac{d^{s+3-r_k}}{d\alpha^{s+3-r_k}} g_k(\alpha).$$

Lemma

Let g_k be the restriction of $g(\alpha)$ to the kth interval I_k, and $r_k = r(\alpha)$ when $\alpha \in I_k$. For any adaptive sampling strategy

$$\frac{d^{s+3}}{d\alpha^{s+3}} g_k(\alpha) = 2\frac{d^{s+3}}{d\alpha^{s+3}} f_k(\alpha) + \frac{(-1)^{s+1-r_k}\cdot s!}{\alpha^{s+1-r_k}(r_k-1)!}\frac{d^{r_k+2}}{d\alpha^{r_k+2}} g_k(\alpha) + \frac{s!}{(1-\alpha)^{r_k}(s-r_k)!}\frac{d^{s+3-r_k}}{d\alpha^{s+3-r_k}} g_k(\alpha).$$

Lemma

For any adaptive sampling strategy

$$\lim_{\alpha \to 0} v(\alpha) = \frac{r_0(s+1)}{(s+1-r_0)((s+2)(s+1)-r_0(r_0+1))},$$

where $r_0 = \lim_{\alpha \to 0} r(\alpha)$.

Example

- Median-of-$(2t+1)$: $v(0) = v(1) = \frac{2}{3t+4}$
- Proportion-from-s: $v(0) = v(1) = \frac{s+1}{s^2(s+3)} \sim \frac{1}{s^2} + \mathcal{O}(s^{-3})$

Lemma

For any adaptive sampling strategy

$$\lim_{\alpha\to 0} v(\alpha) = \frac{r_0(s+1)}{(s+1-r_0)((s+2)(s+1)-r_0(r_0+1))},$$

where $r_0 = \lim_{\alpha\to 0} r(\alpha)$.

Example

- Median-of-$(2t+1)$: $v(0) = v(1) = \frac{2}{3t+4}$
- Proportion-from-s: $v(0) = v(1) = \frac{s+1}{s^2(s+3)} \sim \frac{1}{s^2} + \mathcal{O}(s^{-3})$

The differential equation to find the expectation characteristic function is

$$\frac{d^2\phi}{d\alpha^2} = 6\left(\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right)\phi(\alpha)$$

with $\phi(\alpha) = f'''(\alpha)$

For the second moment characteristic function $g(\alpha)$ we have

$$\frac{d^2\phi}{d\alpha^2} = 6\left(\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right)\phi(\alpha)$$

with $\phi(\alpha) = g^{(iv)}(\alpha)$

The independent term in the ODE for $g(\alpha)$ vanishes, since $f(\alpha) = 2 + 3\alpha(1-\alpha)$ and $f^{(vi)}(\alpha) = 0$.

For the second moment characteristic function $g(\alpha)$ we have

$$\frac{d^2\phi}{d\alpha^2} = 6\left(\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right)\phi(\alpha)$$

That's exactly the same ODE as for $f(\alpha)$!!

with $\phi(\alpha) = g^{(iv)}(\alpha)$

The independent term in the ODE for $g(\alpha)$ vanishes, since $f(\alpha) = 2 + 3\alpha(1-\alpha)$ and $f^{(vi)}(\alpha) = 0$.

- We integrate four times the solution found
- We plug the general form back into the integral equation to determine the value of the arbitrary constants; we also use the symmetry of $g(\alpha)$
- The final solution is

$$
\begin{aligned}
g(\alpha) = &-\frac{288}{35}\alpha^2(\ln(\alpha)+\ln(1-\alpha)) - \frac{288}{35}\ln(1-\alpha) \\
&+\frac{576}{35}\alpha\ln(1-\alpha) + \frac{30}{7} - \frac{24}{245}\alpha^8 + \frac{96}{245}\alpha^7 \\
&-\frac{48}{175}\alpha^6 - \frac{96}{175}\alpha^5 - \frac{48}{35}\alpha^4 + \frac{144}{35}\alpha^3 \\
&-\frac{7332}{1225}\alpha^2 + \frac{132}{35}\alpha,
\end{aligned}
$$

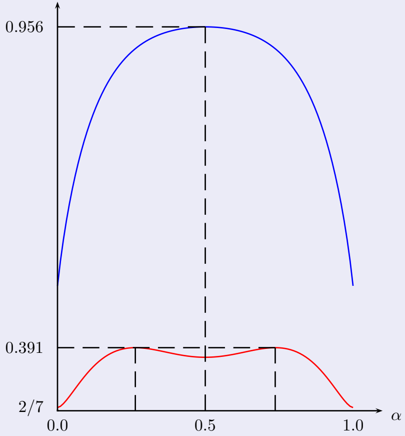
- We integrate four times the solution found
- We plug the general form back into the integral equation to determine the value of the arbitrary constants; we also use the symmetry of $g(\alpha)$
- The final solution is

$$g(\alpha) = -\frac{288}{35}\alpha^2(\ln(\alpha)+\ln(1-\alpha)) - \frac{288}{35}\ln(1-\alpha) + \frac{576}{35}\alpha\ln(1-\alpha) + \frac{30}{7} - \frac{24}{245}\alpha^8 + \frac{96}{245}\alpha^7 - \frac{48}{175}\alpha^6 - \frac{96}{175}\alpha^5 - \frac{48}{35}\alpha^4 + \frac{144}{35}\alpha^3 - \frac{7332}{1225}\alpha^2 + \frac{132}{35}\alpha,$$

- We integrate four times the solution found
- We plug the general form back into the integral equation to determine the value of the arbitrary constants; we also use the symmetry of $g(\alpha)$
- The final solution is

$$
\begin{aligned}
g(\alpha) = &-\frac{288}{35}\alpha^2(\ln(\alpha) + \ln(1-\alpha)) - \frac{288}{35}\ln(1-\alpha) \\
&+ \frac{576}{35}\alpha\ln(1-\alpha) + \frac{30}{7} - \frac{24}{245}\alpha^8 + \frac{96}{245}\alpha^7 \\
&- \frac{48}{175}\alpha^6 - \frac{96}{175}\alpha^5 - \frac{48}{35}\alpha^4 + \frac{144}{35}\alpha^3 \\
&- \frac{7332}{1225}\alpha^2 + \frac{132}{35}\alpha,
\end{aligned}
$$

A plot of $v(\alpha)$ for standard quickselect (Kirschenhofer, Prodinger (1998)) and for median-of-three



We've got the general form of $g(\alpha)$ for standard quickselect and proportion-from-2, but the process of determining the arbitrary constants is still not finished ...

It's much harder than we though!!

- 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 $\alpha > 1/2$)

- 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 $\alpha < 1/2$,

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

The proof of Martínez, Panario, Viola (2004) for adaptive optimal sampling works also for $s = s(n)$, as long as $s \to \infty$and $s/n \to 0$ if $n \to \infty$.

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

Theorem

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

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

Theorem

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

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

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

Theorem

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

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

Theorem

The variance and the expected value of proportion-from-s, with variable-sized samples, is minimized when

$$s = \Theta(\sqrt{n})$$

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

However, the algorithm is more complicated and uses samples of size $\Theta(n^{2/3}\log n)$ (why!?)

Current work:

- Exact solutions for particular strategies (e.g., proportion-from-2)
- Precise asymptotic estimates of the optimal sample size when $s \to \infty$
- We need better estimates of the behavior when $s \to \infty$, e.g., we know that $f(\alpha) = 1 + \min(\alpha, 1-\alpha) + \mathcal{O}(s^{-1})$, but a precise estimate of the s^{-1} term would allow us to compute the factor for the optimal sample size

Current work:

- Exact solutions for particular strategies (e.g., proportion-from-2)
- Precise asymptotic estimates of the optimal sample size when $s \to \infty$
- We need better estimates of the behavior when $s \to \infty$, e.g., we know that $f(\alpha) = 1 + \min(\alpha, 1 - \alpha) + \mathcal{O}(s^{-1})$, but a precise estimate of the s^{-1} term would allow us to compute the factor for the optimal sample size

Current work:

- Exact solutions for particular strategies (e.g., proportion-from-2)
- Precise asymptotic estimates of the optimal sample size when $s \to \infty$
- We need better estimates of the behavior when $s \to \infty$, e.g., we know that $f(\alpha) = 1 + \min(\alpha, 1-\alpha) + \mathcal{O}(s^{-1})$, but a precise estimate of the s^{-1} term would allow us to compute the factor for the optimal sample size