Optimal Sampling for Sorting and Selection

Conrado Martínez

Univ. Politècnica de Catalunya, Spain

February, 2006
1. Introduction

2. Fixed Size Samples

3. Optimal Sampling
QuickSort and quickselect were invented in the early sixties by C.A.R. Hoare (Hoare, 1961; Hoare, 1962).

They are simple, elegant, beautiful and practical solutions to two basic problems of Computer Science: sorting and selection.

They are primary examples of the divide-and-conquer principle.
• Quicksort and quickselect were invented in the early sixties by C.A.R. Hoare (Hoare, 1961; Hoare, 1962)

• They are simple, elegant, beautiful and practical solutions to two basic problems of Computer Science: sorting and selection

• They are primary examples of the divide-and-conquer principle
Quicksort and quickselect were invented in the early sixties by C.A.R. Hoare (Hoare, 1961; Hoare, 1962).

They are simple, elegant, beautiful and practical solutions to two basic problems of Computer Science: sorting and selection.

They are primary examples of the divide-and-conquer principle.
```cpp
void quicksort(vector<Elem>& A, int i, int j) {
    if (i < j) {
        int p = select_pivot(A, i, j);
        swap(A[p], A[1]);
        int k;
        partition(A, i, j, k);
        quicksort(A, i, k - 1);
        quicksort(A, k + 1, j);
    }
}
```
Quickselect

```cpp
Elem quickselect(vector<Elem>& A, 
                 int i, int j, int m) {
    if (i >= j) return A[i];
    int p = select_pivot(A, i, j, m);
    swap(A[p], A[l]);
    int k;
    partition(A, i, j, k);
    if (m < k) quickselect(A, i, k - 1, m);
    else if (m > k) quickselect(A, k + 1, j, m);
    else return A[k];
}
```
void partition(vector<Elem>& A, 
              int i, int j, int& k) {
    int l = i; int u = j + 1; Elem pv = A[i];
    for ( ; ; ) {
        do ++l; while(A[l] < pv);
        do --u; while(A[u] > pv);
        if (l >= u) break;
        swap(A[l], A[u]);
    }
    swap(A[i], A[u]); k = u;
The Recurrences for Average Cost

- Probability that the selected pivot is the $k$-th of $n$ elements: $\pi_{n,k}$
- Average number of comparisons $Q_n$ to sort $n$ elements:

\[
Q_n = n - 1 + \sum_{k=1}^{n} \pi_{n,k} \cdot (Q_{k-1} + Q_{n-k})
\]
The Recurrences for Average Cost

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

$$
C_{n,m} = n - 1 + \sum_{k=m+1}^{n} \pi_{n,k} \cdot C_{k-1,m} \\
+ \sum_{k=1}^{m-1} \pi_{n,k} \cdot C_{n-k,m-k}
$$
Quicksort: The Average Cost

• For the standard variant, the splitting probabilities are $\pi_{n,k} = 1/n$

• Average number of comparisons $Q_n$ to sort $n$ elements (Hoare, 1962):

$$Q_n = 2(n + 1)H_n - 4n$$

$$= 2n \ln n + (2\gamma - 4)n + 2\ln n + O(1)$$

where $H_n = \sum_{1 \leq k \leq n} 1/k = \ln n + O(1)$ is the $n$-th harmonic number.
Quicksort: The Average Cost

- For the standard variant, the splitting probabilities are $\pi_{n,k} = 1/n$
- Average number of comparisons $Q_n$ to sort $n$ elements (Hoare, 1962):

$$Q_n = 2(n + 1)H_n - 4n = 2n \ln n + (2\gamma - 4)n + 2\ln n + O(1)$$

where $H_n = \sum_{1 \leq k \leq n} 1/k = \ln n + O(1)$ is the $n$-th harmonic number.
Quickselect: The Average Cost

- Average number of comparisons $C_{n,m}$ to select the $m$-th out of $n$ elements (Knuth, 1971):

$$C_{n,m} = 2(n + 3 + (n + 1)H_n - (n + 3 - m)H_{n+1-m} - (m + 2)H_m).$$

- This is $\Theta(n)$ for any $m, 1 \leq m \leq n$. 
Quickselect: The Average Cost

- Average number of comparisons $C_{n,m}$ to select the $m$-th out of $n$ elements (Knuth, 1971):

$$C_{n,m} = 2(n + 3 + (n + 1)H_n - (n + 3 - m)H_{n+1-m} - (m + 2)H_m).$$

- This is $\Theta(n)$ for any $m, 1 \leq m \leq n$. 
Quickselect: The Average Cost

The expectation characteristic function

\[ m_0(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n} = 2 + 2 \cdot \mathcal{H}(\alpha), \]

\[ \mathcal{H}(x) = -(x \ln x + (1 - x) \ln(1 - x)). \]

with \( 0 \leq \alpha \leq 1. \)

- The maximum is at \( \alpha = 1/2, \) where
  
  \[ m_0(1/2) = 2 + 2 \ln 2 = 3.386 \ldots \]

- The mean value is \( \overline{m}_0 = 3 \implies \) the average number of comparisons to select an item of given random rank is \( 3n + o(n). \)
Quickselect: The Average Cost

The expectation characteristic function

\[ m_0(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n} = 2 + 2 \cdot \mathcal{H}(\alpha), \]

\[ \mathcal{H}(x) = -(x \ln x + (1 - x) \ln(1 - x)). \]

with \( 0 \leq \alpha \leq 1. \)

The maximum is at \( \alpha = 1/2, \) where

\[ m_0(1/2) = 2 + 2 \ln 2 = 3.386 \ldots \]

The mean value is \( \bar{m}_0 = 3 \implies \) the average number of comparisons to select an item of given random rank is \( 3n + o(n). \)
Quickselect: The Average Cost

- The expectation characteristic function

\[ m_0(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n} = 2 + 2 \cdot \mathcal{H}(\alpha), \]

\[ \mathcal{H}(x) = -(x \ln x + (1 - x) \ln(1 - x)). \]

with \( 0 \leq \alpha \leq 1 \).

- The maximum is at \( \alpha = 1/2 \), where

\[ m_0(1/2) = 2 + 2 \ln 2 = 3.386 \ldots \]

- The mean value is \( \bar{m} = 3 \implies \) the average number of comparisons to select an item of given random rank is \( 3n + o(n) \).
Variance and More

- The variance of both quicksort and quickselect is $\Theta(n^2)$ (Hennequin, 1989; Kirschenhofer & Prodinger, 1998) implies concentration around the mean for quicksort, not for quickselect.

- Higher moments are also known (e.g., Hennequin, 1989).

- Many properties about the distributions are known (e.g. Régnier, 1989, Rösler, 1991, McDiarmid & Hayward, 1996), but no closed form.
Variance and More

- The variance of both quicksort and quickselect is $\Theta(n^2)$ (Hennequin, 1989; Kirschenhofer & Prodinger, 1998) $\Rightarrow$ concentration around the mean for quicksort, not for quickselect.
- Higher moments are also known (e.g., Hennequin, 1989).
- Many properties about the distributions are known (e.g., Régnier, 1989, Rösler, 1991, McDiarmid & Hayward, 1996), but no closed form.
Variance and More

- The variance of both quicksort and quickselect is $\Theta(n^2)$ (Hennequin, 1989; Kirschenhofer & Prodinger, 1998) $\Rightarrow$ concentration around the mean for quicksort, not for quickselect.

- Higher moments are also known (e.g., Hennequin, 1989).

- Many properties about the distributions are known (e.g. Régnier, 1989, Rösler, 1991, McDiarmid & Hayward, 1996), but no closed form.
Improving Quicksort and Quickselect

- Apply general techniques: recursion removal, loop unwrapping, ...
- Reorder recursive calls to quicksort
- Switch to a simpler algorithm for small subfiles
- Use samples to select better pivots
Improving Quicksort and Quickselect

- Apply general techniques: recursion removal, loop unwrapping, ...
- Reorder recursive calls to quicksort
- Switch to a simpler algorithm for small subfiles
- Use samples to select better pivots
Improving Quicksort and Quickselect

- Apply general techniques: recursion removal, loop unwrapping, ...
- Reorder recursive calls to quicksort
- Switch to a simpler algorithm for small subfiles
- Use samples to select better pivots
Improving Quicksort and Quickselect

- Apply general techniques: recursion removal, loop unwrapping, ...
- Reorder recursive calls to quicksort
- Switch to a simpler algorithm for small subfiles
- Use samples to select better pivots
Improving Quicksort and Quickselect

- Apply general techniques: recursion removal, loop unwrapping, ...
- Reorder recursive calls to quicksort
- Switch to a simpler algorithm for small subfiles
- Use samples to select better pivots
1. Introduction

2. Fixed Size Samples

3. Optimal Sampling
Quicksort with Median-of-Three

- In quicksort with median-of-three, the pivot of each recursive stage is selected as the median of a sample of three elements (Singleton, 1969)
- This reduces the probability of uneven partitions which lead to quadratic worst-case
Quicksort with Median-of-Three

- In quicksort with median-of-three, the pivot of each recursive stage is selected as the median of a sample of three elements (Singleton, 1969)
- This reduces the probability of uneven partitions which lead to quadratic worst-case
Quicksort with Median-of-Three

- The splitting probabilities are

\[ \pi_{n,k} = \frac{(k-1)(n-k)}{\binom{n}{3}} \]

- The average number of comparisons made by quicksort with median-of-three \( Q_n \) is (Sedgewick, 1975)

\[ Q_n = \frac{12}{7} n \log n + O(n), \]

roughly a 14.3% less than standard quicksort.
Quicksort with Median-of-Three

- The splitting probabilities are

\[ \pi_{n,k} = \frac{(k-1)(n-k)}{\binom{n}{3}} \]

- The average number of comparisons made by quicksort with median-of-three \( Q_n \) is (Sedgewick, 1975)

\[ Q_n = \frac{12}{7} n \log n + O(n), \]

roughly a 14.3% less than standard quicksort
Quickselect with Median-of-Three

- The average number of comparisons $C_{n,m}$ made by quickselect with median-of-three is (Kirschenhofer, Martínez & Prodinger, 1997)

$$C_{n,m} = 2n + \frac{72}{35}H_n - \frac{156}{35}H_m - \frac{156}{35}H_{n+1-m} + 3m - \frac{(m-1)(m-2)}{n} + O(1)$$

- To obtain this result we used the bivariate generating function

$$C(z, u) = \sum_{n \geq 0} \sum_{1 \leq m \leq n} C_{n,m} z^n u^m$$
Quickselect with Median-of-Three

- The average number of comparisons $C_{n,m}$ made by quickselect with median-of-three is (Kirschenhofer, Martínez & Prodinger, 1997)

$$C_{n,m} = 2n + \frac{72}{35} H_n - \frac{156}{35} H_m - \frac{156}{35} H_{n+1-m}$$

$$+ 3m - \frac{(m - 1)(m - 2)}{n} + O(1)$$

- To obtain this result we used the bivariate generating function

$$C(z, u) = \sum_{n \geq 0} \sum_{1 \leq m \leq n} C_{n,m} z^n u^m$$
Quickselect with Median-of-Three

- The recurrences translate into a second-order differential equation of hypergeometric type satisfied by $C(z, u)$
- We compute then explicit solutions for the GF, and from there, one has to extract (painfully) the coefficients
Quickselect with Median-of-Three

- The recurrences translate into a second-order differential equation of hypergeometric type satisfied by $C(z, u)$
- We compute then explicit solutions for the GF, and from there, one has to extract (painfully 😞) the coefficients
Quickselect with Median-of-Three

- The expectation characteristic function is

\[ m_1(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n} = 2 + 3 \cdot \alpha \cdot (1 - \alpha) \]

with \( 0 \leq \alpha \leq 1 \).

- For any \( \alpha \), \( m_1(\alpha) \leq m_0(\alpha) \)

- The mean value is \( \overline{m}_1 = 5/2 \); compare to \( 3n + o(n) \) comparisons for standard quickselect on random ranks.
Quickselect with Median-of-Three

- The expectation characteristic function is

\[
m_1(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n} = 2 + 3 \cdot \alpha \cdot (1 - \alpha)
\]

with \(0 \leq \alpha \leq 1\).

- For any \(\alpha\), \(m_1(\alpha) \leq m_0(\alpha)\)

- The mean value is \(m_1 = 5/2\); compare to \(3n + o(n)\) comparisons for standard quickselect on random ranks

Optimal Sampling for Sorting and Selection

INCO, Uruguay
Quickselect with Median-of-Three

- The expectation characteristic function is
  \[ m_1(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n} = 2 + 3 \cdot \alpha \cdot (1 - \alpha) \]
  with \(0 \leq \alpha \leq 1\).

- For any \(\alpha\), \(m_1(\alpha) \leq m_0(\alpha)\)

- The mean value is \(\bar{m}_1 = 5/2\); compare to \(3n + o(n)\) comparisons for standard quickselect on random ranks.
Quickselect with Median-of-Three

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

\[ m_0(\alpha) \]

\[ m_1(\alpha) \]

\[ 3.386 \ldots \]

\[ 2.75 \]

\[ 2 \]

\[ 0.0 \quad 0.5 \quad 1.0 \]

\[ \alpha \]
Median-of-\((2t + 1)\)

- The generalization to samples of size \(s = (2t + 1)\) is immediate.
- If \(s = \Theta(1)\) then the recurrences for quicksort and quickselect are \(\sim\) as for the standard case \((s = 1)\).
- The splitting probabilities are:

\[
\pi_{n,k} = \frac{(k-1)(n-k)}{\binom{n}{2t+1}}
\]
Median-of-$(2t+1)$

- The generalization to samples of size $s = (2t+1)$ is immediate.
- If $s = \Theta(1)$ then the recurrences for quicksort and quickselect are $\sim$ as for the standard case ($s = 1$).
- The splitting probabilities are:

\[
\pi_{n,k} = \frac{\binom{k-1}{t} \binom{n-k}{t}}{\binom{n}{t+1}}
\]
**Median-of-\((2t+1)\)**

- The generalization to samples of size \(s = (2t+1)\) is immediate.
- If \(s = \Theta(1)\) then the recurrences for quicksort and quickselect are \(~\) as for the standard case \((s = 1)\).
- The splitting probabilities are:

\[
\pi_{n,k} = \frac{(k-1) \binom{n-k}{t}}{\binom{n}{2t+1}}
\]
Quicksort with Median-of-\((2t+1)\)

- Average number of comparisons \(Q_n^{(t)}\) (VanEmden, 1970)
  \[
  Q_n^{(t)} = \frac{1}{H_{2t+2} - H_{t+1}} n \log n + \mathcal{O}(n)
  \]

- Notice that \(c_t = 1/(H_{2t+2} - H_{t+1})\) tends to \(1/\ln 2\) as \(t \to \infty\); this means that with large samples
  \[Q_n \sim n \log_2 n\]

which is optimal (in the theoretical sense)
**Quicksort with Median-of-\((2t+1)\)**

- **Average number of comparisons** \(Q_n^{(t)}\) (VanEmden, 1970)

\[
Q_n^{(t)} = \frac{1}{H_{2t+2} - H_{t+1}} n \log n + \mathcal{O}(n)
\]

- **Notice** that \(c_t = 1/(H_{2t+2} - H_{t+1})\) tends to \(1/\ln 2\) as \(t \to \infty\); this means that with large samples

\[
Q_n \sim n \log_2 n
\]

which is optimal (in the theoretical sense)
Quickselect with Median-of-\((2t + 1)\)

- The average number of comparisons is not known; must be linear, but the coefficient \(m_t(\alpha)\) remains unknown.

- Average number of comparisons \(C_{n}^{(t)}\) to select an element of random rank (Martínez & Roura, 2001):
  \[
  C_{n}^{(t)} = (2 + \frac{1}{t + 1})n + o(n)
  \]

- The variance of the number of comparisons to select an element of random rank (Martínez & Roura, 2001):
  \[
  \text{Var}\left[C_{n}^{(t)}\right] = \frac{2t + 3}{3(t + 1)^{2}}n^2 + o(n^2)
  \]
Quickselect with Median-of-(2t + 1)

- The average number of comparisons is not known; must be linear, but the coefficient $m_t(\alpha)$ remains unknown.
- Average number of comparisons $C_n^{(t)}$ to select an element of random rank (Martínez & Roura, 2001):
  \[ C_n^{(t)} = (2 + \frac{1}{t + 1})n + o(n) \]

- The variance of the number of comparisons to select an element of random rank (Martínez & Roura, 2001):
  \[ \sqrt{\text{Var}[C_n^{(t)}]} = \frac{2t + 3}{3(t + 1)^2} n^2 + o(n^2) \]
Quickselect with Median-of-$(2t + 1)$

- The average number of comparisons is not known; must be linear, but the coefficient $m_t(a)$ remains unknown.

- Average number of comparisons $C_n^{(t)}$ to select an element of random rank (Martínez & Roura, 2001):
  
  $$C_n^{(t)} = (2 + \frac{1}{t+1})n + o(n)$$

- The variance of the number of comparisons to select an element of random rank (Martínez & Roura, 2001):
  
  $$\nabla \left[ C_n^{(t)} \right] = \frac{2t + 3}{3(t+1)^2} n^2 + o(n^2)$$
Median-of-\((2t + 1)\)

- The main technique to obtain the results was the continuos master theorem (Roura, 1997); it allows to solve many recurrences of the type

\[
F_n = t_n + \sum_{0 \leq k < n} \omega_{n,k} F_k
\]

- The CMT is a powerful generalization of the usual master theorem found in textbooks (e.g., Cormen, Leiserson & Rivest, 1990)
Median-of-\((2t + 1)\)

- The main technique to obtain the results was the continuous master theorem (Roura, 1997); it allows to solve many recurrences of the type

\[
F_n = t_n + \sum_{0 \leq k < n} \omega_{n,k} F_k
\]

- The CMT is a powerful generalization of the usual master theorem found in textbooks (e.g., Cormen, Leiserson & Rivest, 1990)
Median-of-\((2t + 1)\)

- To use the CMT one needs to find a continuous approximation of the weights \(w_{n,k}\); we typically use \(\omega(z) = \lim_{n \to \infty} n \cdot w_{n,z,n}\).
- Then one has to compute

\[
\mathcal{H} = 1 - \int_0^1 \omega(z) \cdot z^a \, dz
\]

where \(a > -1\) is the exponent of \(n\) in \(t_n\); we have three cases depending on \(\mathcal{H} > 0\), \(\mathcal{H} = 0\), \(\mathcal{H} < 0\).
Median-of-$(2t + 1)$

To use the CMT one needs to find a continuous approximation of the weights $w_{n,k}$; we typically use $w(z) = \lim_{n \to \infty} n \cdot w_{n,z,n}$

Then one has to compute

$$H = 1 - \int_0^1 w(z) \cdot z^a \, dz$$

where $a > -1$ is the exponent of $n$ in $t_n$; we have three cases depending on $H > 0$, $H = 0$, $H < 0$
Adaptive Sampling for Quickselect

- Median-of-$(2t + 1)$ might be a good idea for sorting: both subarrays must be recursively sorted; but it is not so natural for selection.

- In proportion-from-$s$ sampling we take an element in the sample of $s$ elements whose rank is, in relative terms, close to the rank of the sought element (Martínez, Panario & Viola, 2004).
Adaptive Sampling for Quickselect

- Median-of-$(2t + 1)$ might be a good idea for sorting: both subarrays must be recursively sorted; but it is not so natural for selection.

- In proportion-from-$s$ sampling we take an element in the sample of $s$ elements whose rank is, in relative terms, close to the rank of the sought element (Martínez, Panario & Viola, 2004).
Adaptive Sampling for Quickselect

- More generally, if the current relative rank is $\alpha = m/n$, we select the element of rank $r(\alpha)$ from the sample as our pivot.

**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 for Quickselect

More generally, if the current relative rank is $\alpha = m/n$, we select the element of rank $r(\alpha)$ from the sample as our pivot.

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 for Quickselect

Example

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

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

\[
\alpha = \frac{4}{15} < \frac{1}{3}
\]
Adaptive Sampling for Quickselect

Example

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

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

\(\alpha = 4/15 < 1/3\)
Adaptive Sampling for Quickselect

Example

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

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

\(\alpha = 4/15 < 1/3\)
Adaptive Sampling for Quickselect

Example

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

Example

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

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

\[
1/3 < \alpha = 4/8 = 1/2 < 2/3
\]
Adaptive Sampling for Quickselect

Example

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

\[
\begin{array}{cccccccc}
7 & 5 & 4 & 6 & 3 & 1 & 8 & 2 \rule{0pt}{2.5ex} \textcolor{black}{9} \rule{0pt}{1.5ex} & 15 & 11 & 13 & 12 & 10 & 14
\end{array}
\]

\[
1/3 < \alpha = 4/8 = 1/2 < 2/3
\]
Adaptive Sampling for Quickselect

Example

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

Example

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

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

\[\alpha = 4/5 > 2/3\]
Adaptive Sampling for Quickselect

**Example**

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

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

\[\alpha = \frac{4}{5} > \frac{2}{3}\]
Adaptive Sampling for Quickselect

Example

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

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

Theorem (Martínez, Panario & Viola, 2004)

For any adaptive sampling strategy, the expectation characteristic function \( f(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}}{n} \) satisfies

\[
f(\alpha) = 1 + \frac{s!}{(r(\alpha) - 1)!(s - r(\alpha))!} \times \left[ \int_{\alpha}^{1} f\left( \frac{\alpha}{x} \right) x^{r(\alpha)}(1 - x)^{s-r(\alpha)} \, dx \\
+ \int_{0}^{\alpha} f\left( \frac{\alpha - x}{1-x} \right) x^{r(\alpha) - 1}(1 - x)^{s+1-r(\alpha)} \, dx \right]
\]
Adaptive Sampling for Quickselect

**Theorem (Martínez & Daligault, 2006)**

The second factorial moment characteristic function $g(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}(C_{n,m-1})}{n^2}$ of any adaptive sampling strategy satisfies

$$g(\alpha) = 2f(\alpha) - 1$$

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

$$+ \int_{0}^{\alpha} g \left( \frac{\alpha - x}{1 - x} \right) x^{r(\alpha)-1}(1 - x)^{s+2-r(\alpha)} \, dx$$
Adaptive Sampling for Quickselect

Theorem (Martínez & Daligault, 2006)

The second factorial moment characteristic function
\[ g(\alpha) = \lim_{n \to \infty, m/n \to \alpha} \frac{C_{n,m}(C_{n,m-1})}{n^2} \]

of any adaptive sampling strategy satisfies

\[ g(\alpha) = 2f(\alpha) - 1 \]

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

\[ + \int_{0}^{\alpha} g \left( \frac{\alpha - x}{1 - x} \right) x^{r(\alpha)-1}(1 - x)^{s+2-r(\alpha)} \, dx \]
Adaptive Sampling for Quickselect

A plot of median-of-three characteristic function versus proportion-from-three $f(\alpha)$.
Adaptive Sampling for Quickselect

A plot of $v(\alpha)$ for standard quickselect (Kirschenhofer & Prodinger, 1998) and for median-of-three (Martínez & Daligault, 2006).
Adaptive Sampling for Quickselect

- With a suitable choice of the endpoints of the intervals that define $r(\alpha)$, we have shown that there exists a proportion-from-3-like strategy which makes the minimum average number of comparisons for all $\alpha$ (among all strategies using samples of three elements).

- The same techniques can be used to find the strategy which minimizes the average total cost (a weighted sum of exchanges and comparisons).
Adaptive Sampling for Quickselect

- With a suitable choice of the endpoints of the intervals that define $r(\alpha)$, we have shown that there exists a proportion-from-3-like strategy which makes the minimum average number of comparisons for all $\alpha$ (among all strategies using samples of three elements).

- The same techniques can be used to find the strategy which minimizes the average total cost (a weighted sum of exchanges and comparisons).
1. Introduction

2. Fixed Size Samples

3. Optimal Sampling
Optimal Sampling for Quicksort

- We consider now samples of size $s = s(n) = 2t(n) + 1$, with $t = o(n)$ and $t \to \infty$ as $n \to \infty$, for instance $t = \log n$.

- The recurrence for the average cost is now:

$$Q_n = n + \Theta(s) + \sum_{k=1}^{n} \pi_{n,k} \cdot (Q_{k-1} + Q_{n-k}),$$

It's important to take into account the work done to select the pivot from the sample!
Optimal Sampling for Quicksort

- We consider now samples of size 
  \[ s = s(n) = 2t(n) + 1, \text{ with } t = o(n) \text{ and } t \to \infty \text{ as } n \to \infty, \text{ for instance } t = \log n \]

- The recurrence for the average cost is now

\[
Q_n = n + \Theta(s) + \sum_{k=1}^{n} \pi_{n,k} \cdot (Q_{k-1} + Q_{n-k}),
\]

it's important to take into account the work done to select the pivot from the sample!
Optimal Sampling for Quicksort

- The standard techniques for fixed-size samples do not work here, the basic problem are the splitting probabilities $\pi_{n,k}$
- The CMT comes to rescue to allow us rigorously prove "handwaving" intuitive arguments ...
Optimal Sampling for Quicksort

- The standard techniques for fixed-size samples do not work here, the basic problem are the splitting probabilities $\pi_{n,k}$.
- The CMT comes to rescue to allow us rigorously prove "handwaving" intuitive arguments …
Optimal Sampling for Quicksort

Theorem (Martínez & Roura, 2001)

The average number of comparisons made by quicksort with median-of-\((2t + 1)\), for \(t = t(n)\) satisfying \(t \to \infty\) and \(t/n \to 0\) when \(n \to \infty\), is

\[
Q_n = n \log_2 n + o(n \log n)
\]
Optimal Sampling for Quicksort

Theorem (Martínez & Roura, 2001)

The average total cost (\# comparisons + \(\xi\) \cdot \# exchanges) of quicksort with median-of-(2t + 1), for \(t = t(n)\) satisfying \(t \rightarrow \infty\) and \(t/n \rightarrow 0\) when \(n \rightarrow \infty\), is

\[
\hat{Q}_n = (1 + \xi/4) \cdot n \log_2 n + o(n \log n),
\]
Computing the Optimal Sample Size

- The idea is to substitute the asymptotic when $t \to \infty$ into the recurrences

$$Q_n = n + \Theta(s) + \sum_{k=0}^{n-1} \pi_{n,k+1} \cdot \left( k \log_2 k + (n-k) \log_2(n-k) \\
+ o(k \log k + (n-k) \log(n-k)) \right),$$

- ...and compute asymptotic estimates of the right hand-side

$$Q_n = n + \beta \cdot s + \frac{n \log_2 n}{2s} + o(s),$$

where we put $\beta \cdot s + o(s)$ the (average) cost of selecting the median from the sample.
Computing the Optimal Sample Size

- The idea is to substitute the asymptotic when $t \to \infty$ into the recurrences

$$Q_n = n + \Theta(s) + \sum_{k=0}^{n-1} \pi_{n,k+1} \cdot \left( k \log_2 k + (n - k) \log_2(n - k) \right) + o(k \log k + (n - k) \log(n - k)), $$

...and compute asymptotic estimates of the right hand-side

$$Q_n = n + \beta \cdot s + \frac{n \log_2 n}{2s} + o(s),$$

where we put $\beta \cdot s + o(s)$ the (average) cost of selecting the median from the sample.
Optimal Sampling for Quicksort

Theorem (Martínez & Roura, 2001)

Let \( s^* = 2t^* + 1 \) denote the optimal sample size that minimizes the average number of comparisons made by quicksort. Then

\[
t^* = \sqrt{\frac{1}{\beta}} \left( \frac{4 - \xi (2 \ln 2 - 1)}{8 \ln 2} \right) \cdot \sqrt{n} + o\left(\sqrt{n}\right)
\]

if \( \xi < \tau = 4/(2 \ln 2 - 1) \approx 10.3548 \)
Optimal Sample Sizes for Quicksort

Optimal sample size vs. exact values

![Graph showing the relationship between optimal sample size and exact values](image-url)
Expensive Exchanges and Optimal Sampling

- If exchanges are expensive ($\xi \geq \tau$), pick the $(\psi \cdot s)$-th element of a sample of size $\Theta(\sqrt{n})$, not the median.
- If the position of the pivot is close to either end of the array, then very few exchanges are necessary on that stage, but a poor partition leads to more recursive steps. This trade-off is relevant if exchanges are very expensive.
- We found an explicit formula for $\psi$ as a function of $\xi$. 
Expensive Exchanges and Optimal Sampling

- If exchanges are expensive ($\xi \geq \tau$), pick the $(\psi \cdot s)$-th element of a sample of size $\Theta(\sqrt{n})$, not the median.

- If the position of the pivot is close to either end of the array, then very few exchanges are necessary on that stage, but a poor partition leads to more recursive steps. This trade-off is relevant if exchanges are very expensive.

- We found an explicit formula for $\psi$ as a function of $\xi$. 
Expensive Exchanges and Optimal Sampling

- If exchanges are expensive ($\xi \geq \tau$), pick the $(\psi \cdot s)$-th element of a sample of size $\Theta(\sqrt{n})$, not the median.

- If the position of the pivot is close to either end of the array, then very few exchanges are necessary on that stage, but a poor partition leads to more recursive steps. This trade-off is relevant if exchanges are very expensive.

- We found an explicit formula for $\psi$ as a function of $\xi$. 
Optimal Sampling for Quickselect

Theorem (Martínez & Roura, 2001)

The average total cost
(# comparisons + ξ • # exchanges) of quickselect with median-of-(2t + 1) to select an element of random rank, for \( t = t(n) \) satisfying \( t \to \infty \) and \( t/n \to 0 \) when \( n \to \infty \), is

\[
\hat{C}_n = 2(1 + \xi/4) \cdot n + o(n \log n),
\]
Optimal Sampling for Quickselect

Theorem (Martínez & Roura, 2001)

Let \( s^* = 2t^* + 1 \) denote the optimal sample size that minimizes the average total cost of quickselect. Then

\[
t^* = \frac{1}{2\sqrt{\beta}} \cdot \sqrt{n} + o\left(\sqrt{n}\right)
\]
Optimal Sampling for Quickselect

- Solving the integral equations for the expectation and second factorial moment characteristic function is difficult, but we can analyse what happens when $s \to \infty$
- For instance, if we use median-of-$(2t + 1)$ sampling then $m_t(\alpha) = 2$ when $t \to \infty$; this is not optimal
Optimal Sampling for Quickselect

- Solving the integral equations for the expectation and second factorial moment characteristic function is difficult, but we can analyse what happens when $s \to \infty$.

- For instance, if we use median-of-$(2t + 1)$ sampling then $m_t(\alpha) = 2$ when $t \to \infty$; this is not optimal.
Optimal Sampling for Quickselect

Theorem (Martínez, Panario & Viola, 2004)

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

$$f(\alpha) = 1 + \min(\alpha, 1 - \alpha)$$
Optimal Sampling for Quickselect

**Theorem (Martínez & Daligault, 2006)**

The variance of proportion-from-$s$ sampling with $s \to \infty$ is subquadratic. Since

$$g(\alpha) = (1 + \min(\alpha, 1 - \alpha))^2 = f^2(\alpha),$$

we have

$$\lim_{n \to \infty, m/n \to \alpha} \frac{\text{Var}[C_{n,m}]}{n^2} = g(\alpha) - f^2(\alpha) = 0$$
Optimal Sampling for Quickselect

- The two results above hold for biased proportion-from-$s$ strategies.
- The rank $r(\alpha)$ must be close to $\alpha \cdot s$ ... but no too close!
- We want our selected pivot to be close to the sought element, but at the proper side; e.g., if $\alpha < 1/2$ the pivot should be slightly to the right of the sought element, not to the left.
- Solution: take $r(\alpha) > \alpha \cdot s + 1 - \alpha$ if $\alpha < 1/2$ and symmetrically if $\alpha > 1/2.$
Optimal Sampling for Quickselect

- The two results above hold for **biased** proportion-from-s strategies.
- The rank \( r(\alpha) \) must be close to \( \alpha \cdot s \) ... but no too close!
- We want our selected pivot to be close to the sought element, but at the proper side; e.g., if \( \alpha < 1/2 \) the pivot should be slightly to the right of the sought element, not to the left.
- Solution: take \( r(\alpha) > \alpha \cdot s + 1 - \alpha \) if \( \alpha < 1/2 \) and symmetrically if \( \alpha > 1/2 \).
The two results above hold for biased proportion-from-$s$ strategies.

The rank $r(\alpha)$ must be close to $\alpha \cdot s$ ... but no too close!

We want our selected pivot to be close to the sought element, but at the proper side; e.g., if $\alpha < 1/2$ the pivot should be slightly to the right of the sought element, not to the left.

Solution: take $r(\alpha) > \alpha \cdot s + 1 - \alpha$ if $\alpha < 1/2$ and symmetrically if $\alpha > 1/2$.
Optimal Sampling for Quickselect

- The two results above hold for biased proportion-from-$s$ strategies
- The rank $r(\alpha)$ must be close to $\alpha \cdot s$ ... but no too close!
- We want our selected pivot to be close to the sought element, but at the proper side; e.g., if $\alpha < 1/2$ the pivot should be slightly to the right of the sought element, not to the left
- Solution: take $r(\alpha) > \alpha \cdot s + 1 - \alpha$ if $\alpha < 1/2$ and symmetrically if $\alpha > 1/2
Optimal Sampling for Quickselect

- The two results above hold for biased proportion-from-$s$ strategies.
- The rank $r(\alpha)$ must be close to $\alpha \cdot s$ ... but no too close!
- We want our selected pivot to be close to the sought element, but at the proper side; e.g., if $\alpha < 1/2$ the pivot should be slightly to the right of the sought element, not to the left.
- Solution: take $r(\alpha) > \alpha \cdot s + 1 - \alpha$ if $\alpha < 1/2$ and symmetrically if $\alpha > 1/2$. 
Optimal Sampling for Quickselect

- We can plug the asymptotic estimate 
  \[ C_{n,m} = n + \min(m, n - m) + o(n) \] back into quickselect's recurrence to determine the optimal size of samples.

- But it is difficult to obtain precise asymptotics, we only obtained order of magnitude

\[ C_{n,m} = n + \beta \cdot s + \min(m, n - m) + \mathcal{O}\left(\frac{n}{s}\right), \]

\[ \forall [C_{n,m}] = \max\left(n \cdot s, \frac{n^2}{s}\right) \]
Optimal Sampling for Quickselect

- We can plug the asymptotic estimate \( C_{n,m} = n + \min(m, n - m) + o(n) \) back into quickselect's recurrence to determine the optimal size of samples.

- But it is difficult to obtain precise asymptotics, we only obtained order of magnitude

\[
C_{n,m} = n + \beta \cdot s + \min(m, n - m) + \mathcal{O}\left(\frac{n}{s}\right),
\]

\[
\forall[C_{n,m}] = \max\left(n \cdot s, \frac{n^2}{s}\right)
\]
Theorem (Martínez & Daligault, 2006)

Biased proportion-from-$s$ sampling with $s = \Theta(\sqrt{n})$ minimizes both the expectation and variance of the number of comparisons; in particular, the variance is $\Theta(n^{3/2})$. 
Sources

J. Daligault and C. Martínez.
On the variance of quickselect.

P. Kirschenhofer, H. Prodinger, and C. Martínez.
Analysis of Hoare’s Find algorithm with median-of-three partition.
Sources

C. Martínez, D. Panario, and A. Viola.
Adaptive sampling for quickselect.

C. Martínez and S. Roura.
Optimal sampling strategies in quicksort and quickselect.