Old (20th century) and New (21th century) results about Quicksort and Quickselect

Helmut Prodinger

Stellenbosch

November 19, 2008

Comparisons in Quickselect

The traditional sorting method QUICKSORT uses a pivot element and a partitioning strategy in order to bring the pivot element into its correct position and having only smaller elements to the left and larger elements to the right, so that the procedure can be applied recursively to the smaller subfiles.

Comparisons in Quickselect

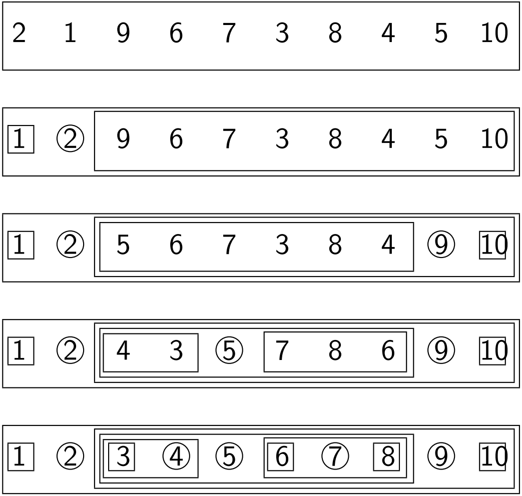
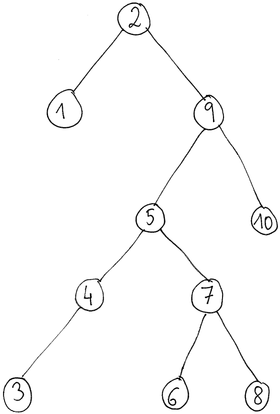


Figure: The partition process

Comparisons in Quickselect



The corresponding binary search tree

Comparisons in Quickselect

If one only wants the find the jth ranked element (which is j, since we assume that the elements in question are $\{1, 2, \ldots, n\}$), one uses the same partitioning strategy, but follows only the path (in the associated binary search tree) which contains the sought element. This is the same as to say that one goes down recursively in only *one* subfile. This procedure is called Hoare's FIND algorithm or QUICKSELECT.

Comparisons in Quickselect

Knuth computed the average number of comparisons $C_{n,j}$. For this, it is assumed that every permutation of $\{1, 2, \dots, n\}$ is equally likely, and that the partitioning phase needs $n-1$ comparisons. Then there is the recursion

$$C_{n,j} = n - 1 + \frac{1}{n} \sum_{1 \le k < j} C_{n-k,j-k} + \frac{1}{n} \sum_{j < k \le n} C_{k-1,j} \ .$$

The solution is

$$C_{n,j} = 2\Big(n + 3 + (n+1)H_n - (j+2)H_j - (n+3-j)H_{n+1-j}\Big) \ .$$

$H_n = \sum_{1 \le k \le n} \frac{1}{k}$ is the nth harmonic number; harmonic numbers of order two: $H_n^{(2)} = \sum_{1 \le k \le n} \frac{1}{k^2}$.

Comparisons in Quickselect

More generally, one can also set up an analogous recursion for the *probability generating functions*, viz.

$$\mathcal{C}_{n,j}(v) = \frac{v^{n-1}}{n}\left[\sum_{1\le k<j} \mathcal{C}_{n-k,j-k}(v) + 1 + \sum_{j<k\le n} \mathcal{C}_{k-1,j}(v)\right].$$

Comparisons in Quickselect

The recursion translates into the fundamental equation

$$\frac{\partial}{\partial z} F(z;u,v) = \frac{F(zv;u,v)}{1-zv} + \frac{u}{(1-zv)(1-zuv)} + \frac{u\,F(zv;u,v)}{1-zuv}\ .$$

Differentiation (and $v = 1$) leads to the generating function of the moments.

Comparisons in Quickselect

The recursion translates into the fundamental equation

$$\frac{\partial}{\partial z} F(z; u, v) = \frac{F(zv; u, v)}{1 - zv} + \frac{u}{(1 - zv)(1 - zuv)} + \frac{u\, F(zv; u, v)}{1 - zuv} \; .$$

Differentiation (and $v = 1$) leads to the generating function of the moments.

Comparisons in Quickselect

One time differentiation:

$$G(z,u) = \frac{2}{(1-u)(1-z)(1-zu)}\left[(u-2)\log\frac{1}{1-zu} + u(2u-1)\log\frac{1}{1-z}\right] + \frac{2zu(z^2u-2zu+3-2z)}{(1-z)^2(1-zu)^2},$$

which gives the coefficients $C_{n,j}$.

Now we differentiate twice:

Comparisons in Quickselect

One time differentiation:

$$G(z,u) = \frac{2}{(1-u)(1-z)(1-zu)}\left[(u-2)\log\frac{1}{1-zu} + u(2u-1)\log\frac{1}{1-z}\right] + \frac{2zu(z^2u-2zu+3-2z)}{(1-z)^2(1-zu)^2},$$

which gives the coefficients $C_{n,j}$.
Now we differentiate twice:

Comparisons in Quickselect

$$
\begin{aligned}
D(z,u) = {} & \frac{4(2-u)^2}{(1-z)(1-zu)(1-u)^2}\log^2\frac{1}{1-zu} \\
& + \frac{4u(1-2u)^2}{(1-z)(1-zu)(1-u)^2}\log^2\frac{1}{1-z} \\
& - \frac{z^2u(-10z^3u^2-5z^2u^2+6zu^2-15u-5z^2u+38zu-15+6z)}{(1-z)^3(1-zu)^3} \\
& + \frac{8u(2-u)(1-2u)}{(1-z)(1-zu)(1-u)^2}\log\frac{1}{1-zu}\log\frac{1}{1-u} \\
& + \frac{8(2-u)(1-2u)}{(1-z)(1-zu)(1-u)^2}\log\frac{1}{1-z}\log\frac{1}{1-u} \\
& + \frac{8u(2-u)(1-2u)}{(1-z)(1-zu)(1-u)^2}\log\frac{1}{1-zu}\log(-u) \\
& - \frac{2u(-5z^2u+19z^2u^2-11zu^2-10zu+z+3u+3)}{(1-u)(1-z)^2(1-zu)^2}\log\frac{1}{1-z} \\
& - \frac{10z^2u^2-38z^2u-2zu^2+20zu+22z-6u-6}{(1-u)(1-z)^2(1-zu)^2}\log\frac{1}{1-zu} \\
& - \frac{8u(2-u)(1-2u)}{(1-z)(1-zu)(1-u)^2}\left(\operatorname{dilog}\left(-\frac{(1-z)u}{1-u}\right)-\operatorname{dilog}\left(-\frac{u}{1-u}\right)\right) \\
& - \frac{8(2-u)(1-2u)}{(1-z)(1-zu)(1-u)^2}\left(\operatorname{dilog}\left(\frac{1-zu}{1-u}\right)-\operatorname{dilog}\left(\frac{1}{1-u}\right)\right).
\end{aligned}
$$

Comparisons in Quickselect

Dilogarithm:

$$\operatorname{dilog} z = \int_1^z \frac{\log t}{1-t}\,dt \ .$$

Comparisons in Quickselect

We read off the coefficient of $u^j z^n$ in $D(z,u)$:

$$
\begin{aligned}
D_{n,j} = {} & -2(n+1)(n+8-2j)H_n^2 - 8(n+1)\,H_n\,H_j \\
& + \frac{(4-32n-12n^2)j^2 + (44n^2+12n^3+28n-4)j + 16n + 16}{j(n+1-j)}\,H_n \\
& + 2(j+8)(j+1)\,H_j^2 \\
& + \Big(-4j^2 + (12+8n)j - 4n^2 + 8 - 12n\Big)\,H_j H_{n+1-j} \\
& - 2\,\frac{(-6n-14)j^3 + (14n+17+6n^2)j^2 + (-3+6n^2+3n)j + 8n + 8}{j(n+1-j)}\,H_j \\
& + \Big(2j^2 + (-4n-18)j + 2n^2 + 18n + 32\Big)\,H_{n+1-j}^2 \\
& - 2\,\frac{(6n+14)j^3 + (-12n^2-25-46n)j^2 + (37n+11+32n^2+6n^3)j + 8n + 8}{j(n+1-j)}\,H_{n+1-j} \\
& + 2(n+1)(n+8-2j)\,H_n^{(2)} - (2j^2+16+10j)\,H_j^{(2)} \\
& + \Big(-2n^2 + 4nj - 18n - 2j^2 - 32 + 18j\Big)\,H_{n+1-j}^{(2)} + \\
& + \frac{10j^4 + (-20n-20)j^3 + (n^2-66-13n)j^2 + (76+9n^3+42n^2+101n)j + 32}{2j(n+1-j)} \\
& + 4(n+2-2j)(n+1)\sum_{k=1}^{j}\frac{H_{n-k}}{k} + 8(n+1-j)\sum_{k=1}^{j}\frac{H_{n+k-j}}{k}\;.
\end{aligned}
$$

Comparisons in Quickselect

Lemma

[Reciprocity law]

$$\sum_{k=1}^{j} \frac{H_{n-k}}{k} + \sum_{k=1}^{n+1-j} \frac{H_{n-k}}{k} = H_j H_{n+1-j} + H_n^2 - H_n^{(2)} - \frac{1}{j(n+1-j)}$$

Comparisons in Quickselect

Lemma

[Reciprocity law]

$$\sum_{k=1}^{j} \frac{H_{n+k-j}}{k} + \sum_{k=1}^{n+1-j} \frac{H_{j+k-1}}{k} = \frac{1}{2}\left(H_j^2 + H_j^{(2)}\right) + \frac{1}{2}\left(H_{n+1-j}^2 + H_{n+1-j}^{(2)}\right) + H_j H_{n+1-j} + \frac{1}{j(n+1-j)} + \frac{n+1}{j(n+1-j)}\left(H_n - H_j - H_{n+1-j}\right).$$

Comparisons in Quickselect

Theorem

The variance of the number of comparisons needed to find the jth element out of n elements is given by the exact formula

$$
\begin{aligned}
\mathrm{Var}_{n,j} = {} & -2(n+1)(3n+8)\,H_n^2 \\
& + 8\,H_n\,H_j\Big(n(j+2)+2\Big) + 8\,H_n\,H_{n+1-j}\Big(n(n+3-j)+2\Big) \\
& + 2\,\frac{-(2n^2+n-9)j^2+(n+1)(2n^2+n-9)j+8(n+1)}{j(n+1-j)}\,H_n \\
& - 2j(j-1)\,H_j^2 - 2(n+1-j)(n-j)\,H_{n+1-j}^2 \\
& + 4\Big(j^2-(n+1)j-n^2-5n-8\Big)\,H_j\,H_{n+1-j} + \qquad \rightarrow \textit{next page}
\end{aligned}
$$

Comparisons in Quickselect

$$
\begin{aligned}
&-\frac{2}{j(n+1-j)}\Big(-(2n+3)j^3+(2n^2+7n+24)j^2-(n+1)(2n+21)j+8(n+1)\Big)\,H_j\\
&-\frac{2}{j(n+1-j)}\Big((2n+3)j^3-(4n^2+8n-15)j^2\\
&\qquad\qquad\qquad+(n+1)(2n^2+3n-18)j+8(n+1)\Big)\,H_{n+1-j}\\
&+2(n+1)(n+6)\,H_n^{(2)}\\
&-2(j^2+5j+8)\,H_j^{(2)}-2(j^2-(2n+7)j+n^2+7n+14)\,H_{n+1-j}^{(2)}\\
&+\frac{10j^4-20(n+1)j^3+(9n^2+31n-6)j^2+(n^2-11n+16)(n+1)j+32}{2j(n+1-j)}\\
&+4n\left((n+1-j)\sum_{k=1}^{j}\frac{H_{n-k}}{k}+j\sum_{k=1}^{n+1-j}\frac{H_{n-k}}{k}\right).\qquad\square
\end{aligned}
$$

Comparisons in Quickselect

In the instance $n = 2N + 1$, $j = N + 1$ (searching for the *median*), we can use the first reciprocity law to simplify the variance. We get

$$
\begin{aligned}
&\mathrm{Var}_{2N+1,N+1} = -8(2N+5)(N+1)H_{2N+1}^2 \\
&+ 16(2N+5)(N+1)H_{2N+1}H_{N+1} \\
&+ 4\frac{4N^3+9N^2+2N+5}{N+1}H_{2N+1} - 8(2N^2+7N+7)H_{N+1}^2 \\
&- 4\frac{4N^3+9N^2-13N-2}{N+1}H_{N+1} + 24(N+1)H_{2N+1}^{(2)} \\
&- 4(N^2+7N+14)H_{N+1}^{(2)} + \frac{7N^4+15N^3+9N^2+5N+20}{(N+1)^2} .
\end{aligned}
$$

Comparisons in Quickselect

Asymptotically we get from this

$$\mathrm{Var}_{2N+1,N+1} \sim N^2\left(16\log 2 - 16\log^2 2 - \frac{2}{3}\pi^2 + 7\right) .$$

The numerical constant is 3.823370396 .

Moves and displacements in Quicksort

New paper in 2008 (with Conrado Martinez)

Moves and displacements of particular elements in Quicksort

Moves and displacements in Quicksort

Two parameters of interest:
1) the number of moves $M_{n,i}$ of element i when we sort an array of size n
2) the (accumulated) displacement $D_{n,i}$ of element i.

Moves and displacements in Quicksort

Analysis of data moves in quicksort involves the so called *quickselect recurrence*.
In its general standard form it reads

$$f_{n,i} = a_{n,i} + \frac{1}{n} \sum_{1 \leq k < i} f_{n-k,i-k} + \frac{1}{n} \sum_{i < k \leq n} f_{k-1,i}, \quad 1 \leq i \leq n,$$

for some given *toll* function $a_{n,i}$.

Moves and displacements in Quicksort

Theorem (Kuba)

The value $f_{n,i}$ defined as above with arbitrary fixed values $a_{n,i}$, $1 \leq i \leq n$, is given by

$$f_{n,i} = a_{1,1} + \sum_{k=n+2-j}^{n} A(k, k-n+j) + \sum_{k=2}^{n+1-j} \frac{ka_{k,1} - (k-1)a_{k-1,1}}{k},$$

where $A(n,i)$ is given by

$$A(n,i) = \sum_{k=j+1}^{n} \frac{ka_{k,i} - (k-1)a_{k-1,i-1} - (k-1)a_{k-1,i} + (k-2)a_{k-2,i-1}}{k} + \frac{ia_{i,i} - (i-1)a_{i-1,i-1}}{i}.$$

Moves and displacements in Quicksort

Theorem

The probability generating function $M_{n,i}(v)$ of the number of moves of element i in a random permutation of $\{1, 2, \ldots, n\}$, satisfies the recursion

$$M_{n,i}(v) = \frac{1}{n} \sum_{1 \le k < i} \left(\frac{k-1}{n-1} v + \frac{n-k}{n-1} \right) M_{n-k,i-k}(v) + \frac{1}{n} \sum_{i < k \le n} \left(\frac{k-2}{n-1} + \frac{n-k+1}{n-1} v \right) M_{k-1,i}(v) + \frac{v}{n}$$

for $n \ge 2$ and $1 \le i \le n$; $M_{1,1}(v) = v$.

Moves and displacements in Quicksort

Corollary

The expected number of moves $\mu_{n,i} = M'_{n,i}(1)$ *satisfies*

$$\mu_{n,i} = \frac{1}{n} \sum_{1 \leq k < i} \mu_{n-k,i-k} + \frac{1}{n} \sum_{i < k \leq n} \mu_{k-1,i} + \frac{1}{n} + \frac{(i-1)(i-2)}{2n(n-1)} + \frac{(n+1-i)(n-i)}{2n(n-1)}, \quad n > 1, 1 \leq i \leq n,$$

with $\mu_{1,1} = 1$.

Moves and displacements in Quicksort

Kuba's theorem with

$$a_{n,i} = \frac{1}{n} + \frac{(i-1)(i-2)}{2n(n-1)} + \frac{(n+1-i)(n-i)}{2n(n-1)}.$$

Theorem

For all $n > 1$, $1 \leq i \leq n$,

$$\mu_{n,i} = \frac{1}{3}H_n + \frac{1}{6}H_i + \frac{1}{6}H_{n+1-i} + \frac{1}{6} + \frac{1}{3i} - \frac{(i-1)^2}{3n} + \frac{(i-1)(i-2)}{3(n-1)} + \frac{1}{12}[\![i=1]\!] - \frac{1}{12}[\![i=n]\!],$$

where $H_n = \sum_{1\leq k\leq n} \frac{1}{k}$ is the n-th harmonic number and $[\![P]\!] = 1$ if P is true and $[\![P]\!] = 0$ otherwise.

Moves and displacements in Quicksort

Corollary

For fixed $i \geq 1$, as $n \to \infty$,

$$\mu_{n,i} = \frac{1}{2}\ln n + \frac{1}{6}H_i + \frac{\gamma}{2} + \frac{1}{3i} + \frac{1}{6} + O(n^{-1}),$$

where $\gamma = 0.577215\ldots$ is Euler's gamma constant.
Furthermore, if $i > 1$,

$$\mu_{n,n+1-i} = \mu_{n,i} - \frac{1}{3i} + O(n^{-1})$$

For $i = \alpha n + o(n)$, $0 < \alpha < 1$, we have

$$\mu_{n,i} = \frac{2}{3}\ln n + \frac{1}{6} + \frac{1}{6}\ln \alpha + \frac{1}{6}\ln(1-\alpha) - \frac{\alpha(1-\alpha)}{3} + \frac{2}{3}\gamma + O(n^{-1}).$$

Moves and displacements in Quicksort

The global minimum of $\mu_{n,i}$ occurs at $i = n$. The maximum of $\mu_{n,\alpha\cdot n}$ occurs close to the median ($\alpha = 1/2$), actually at

$$\alpha^* = \frac{1}{2} - \frac{2}{n} - \frac{39}{n^2} - \frac{582}{n^3} - \frac{8604}{n^4} - \frac{121168}{n^5} + O(n^{-6}),$$

with $\mu_{n,\alpha^* n} = \frac{2}{3}\ln n + \frac{1}{12} - \frac{1}{3}\ln 2 + \frac{2}{3}\gamma + O(n^{-1})$.

Moves and displacements in Quicksort

Another quantity of interest is the cumulated number of moves. By linearity, its expected value is the sum of the $\mu_{n,i}$'s.

Corollary

For $n \geq 2$, the total number of moves is given by

$$\overline{\mu}_n = \sum_{i=1}^{n} \mu_{n,i} = \frac{2}{3}(n+1)H_n - \frac{4n+1}{18}.$$

Moves and displacements in Quicksort

Theorem

The probability generating function $D_{n,i}(v)$ of the displacement of element i in a random permutation of $\{1, 2, \ldots, n\}$, satisfies the recursion

$$D_{n,i}(v) = \frac{1}{n} \sum_{1 \le k < i} \left(\sum_{j=2}^{k} \sum_{\ell=k+1}^{n} \frac{v^{\ell-j}}{(n-1)(n-k)} + \frac{n-k}{n-1} \right) D_{n-k,i-k}(v)$$

$$+ \frac{1}{n} \sum_{i < k \le n} \left(\frac{k-2}{n-1} + \sum_{j=1}^{k-1} \sum_{\ell=k}^{n} \frac{v^{\ell-j}}{(n-1)(k-1)} \right) D_{k-1,i}(v) + \frac{v^{i-1}}{n}$$

for $n \ge 2$ and $1 \le i \le n$; $D_{1,1}(v) = 1$.

Moves and displacements in Quicksort

Corollary

The expected displacement $\delta_{n,i} = D'_{n,i}(1)$ *satisfies*

$$\delta_{n,i} = \frac{1}{n} \sum_{1 \le k < i} \delta_{n-k,i-k} + \frac{1}{n} \sum_{i < k \le n} \delta_{k-1,i} + \frac{(i-1)(i-2)}{4n} + \frac{(n-i)(n+1-i)}{4(n-1)} + \frac{i-1}{n}$$

with $\delta_{1,1} = 0$.

Moves and displacements in Quicksort

Theorem

For all $n > 1$ *and* $1 \leq i \leq n$,

$$\begin{aligned}\delta_{n,i} = {} & \frac{n}{2} + \frac{1}{12}H_n - \frac{1}{12}H_{i-1} - \frac{1}{3}H_{n+1-i} + \frac{5}{24} \\ & - \frac{(i-1)^2}{12n} + \frac{(i-1)(i-2)}{12(n-1)} \\ & + \frac{1}{6}[\![i=1]\!] + \frac{1}{8}[\![i=n]\!].\end{aligned}$$

Moves and displacements in Quicksort

Corollary

For fixed $i \geq 1$, *as* $n \to \infty$,

$$\delta_{n,i} = \frac{n}{2} - \frac{1}{4} \ln n + O(1),$$
$$\delta_{n,n+1-i} = \frac{n}{2} + O(1).$$

For $i = \alpha n + o(n)$, $0 < \alpha < 1$,

$$\delta_{n,i} = \frac{n}{2} - \frac{1}{3} \ln n - \frac{1}{12} \ln \alpha - \frac{1}{3} \ln(1-\alpha) - \frac{\alpha}{12} + \frac{\alpha^2}{12} + \frac{5}{24} + O(n^{-1}).$$

Moves and displacements in Quicksort

The maximum of $\delta_{n,i}$ occurs at $i = n$; there $\delta_{n,n} = n/2$. The minimum of $\delta_{n,\alpha n}$ occurs at

$$\alpha^* = \frac{5}{4} - \frac{\sqrt{17}}{4} + \Big(\frac{5}{8} + \frac{5}{136}\sqrt{17}\Big)\frac{1}{n} + \Big(-\frac{33}{256} + \frac{821}{221952}\sqrt{17}\Big)\frac{1}{n^2}$$
$$+\Big(\frac{981}{4096}+\frac{4864631}{60370944}\sqrt{17}\Big)\frac{1}{n^3}+O(n^{-4}) = 0.219223594\ldots+O(n^{-1}),$$

with

$$\delta_{n,\alpha^* n} = \frac{n}{2} - \frac{1}{3}\ln n - \frac{1}{12}\ln\Big(\frac{5}{4} - \frac{\sqrt{17}}{4}\Big) - \frac{1}{3}\ln\Big(-\frac{1}{4} + \frac{\sqrt{17}}{4}\Big)$$
$$+ \frac{31}{96} - \frac{\sqrt{17}}{32} - \frac{\gamma}{3} + O(n^{-1}).$$