Alcom-FT Technical Report Series
ALCOMFT-TR-03-2

Adaptive Sampling for Quickselect*

Conrado Martínez[1], Daniel Panario[2], Alfredo Viola[3]

[1] Departament de Llenguatges i Sistemes Informàtics,
Universitat Politècnica de Catalunya, E-08034 Barcelona, Spain
E-mail: conrado@lsi.upc.es

[2] School of Mathematics and Statistics, Carleton University
Ottawa, K1S 5B6, Canada
E-mail: daniel@math.carleton.ca

[3] Instituto de Computación, Universidad de la República
Casilla de Correo 16120, Distrito 6, Montevideo, Uruguay
E-mail: viola@fing.edu.uy

Abstract. The median-of-3 variant of quickselect has been largely studied. However, the following natural adaptive variant had not been analyzed: "choose as pivot the smallest of the sample if the rank of the sought element is small, the largest if the rank is large, and the median if the rank is medium". We first analyze proportional-of-2 and proportional-of-3 in which we use equal-size intervals. We propose ν-find, a generalization of median-of-3 and proportional-of-3 with interval breakpoints ν and $1-\nu$. We give the optimal ν and values for which ν-find outperforms median-of-3.

1 Introduction

Hoare's quickselect [5] algorithm selects the mth smallest element (equivalently, the element of *rank* m in ascending order, the mth order statistic) out of an array of n elements by picking an element from the array —the pivot— and rearranging the array so that elements smaller than the pivot are to its left and elements larger than the pivot are to its right. If the pivot has been brought to position $j = m$ then it is the sought element; otherwise, if $m < j$ then the procedure is recursively applied to the subarray to the left of the pivot, and if $m > j$ the process continues in the right subarray. A similar principle is used in the celebrated quicksort algorithm [6], also by Hoare; once the pivot is brought into place by the partitioning of the array, the subarrays to its left and right are recursively sorted.

* The research of the first author was supported by the Future and Emergent Technologies programme of the EU under contract IST-1999-14186 (ALCOM-FT) and DGES PB98-0926 (AEDRI). The second author was funded by NSERC grant number 238757. The third author was supported in part by Proyecto de Investigación CSIC Fondos 2000-2002 and 2002-2004.

Quickselect performs well in practice, its average cost being linear. In particular,

$$m_0(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \frac{C^{(0)}_{n,m}}{n} = 2\cdot(1+\mathcal{H}(\alpha)), \qquad 0\le\alpha\le 1,$$

where $C^{(0)}_{n,m}$ denotes the average number of comparisons made by quickselect to select the mth element out of n, and $\mathcal{H}(x) = -x\ln x-(1-x)\ln(1-x)$ (see [4, 8–10]). Another quantity of interest is the average cost $C^{(0)}_n$ to locate an element of random rank, i.e., when m is given by a uniformly distributed random variable in $\{1,\dots,n\}$; since $C^{(0)}_n = 1/n\cdot\sum_{1\le m\le n} C^{(0)}_{n,m}$ we have $\overline{m}_0 = \lim_{n\to\infty} C^{(0)}_n/n = 3$.

Using the median of a small sample of $s = 2t+1$ elements as the pivot of each recursive stage yields a substantial improvement over the standard variant, reducing the average number of comparisons and making worst-case behavior less likely. Little is known for $t>1$, other than the average and variance of the number of comparisons to select an element of random rank out of n [11]. For samples of three elements [1, 3, 7] it is known that

$$m_1(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \frac{C^{(1)}_{n,m}}{n} = 2+3\alpha(1-\alpha), \qquad 0\le\alpha\le 1,$$

where, for any t, $C^{(t)}_{n,m}$ denotes the average number of comparisons made by quickeselect with median-of-$(2t+1)$ sampling to select the mth out of n. Also, $\overline{m}_1 = \lim_{n\to\infty} C^{(1)}_n/n = 5/2$.

However, median-of-$(2t+1)$ sampling is not a natural pivot selection strategy for quickselect. For instance, if we were looking for the 100th element in a collection of 1000 elements it would seem more natural to pick the smallest element of a sample of three rather than the median. We call these *adaptive* sampling techniques. In general, it seems better and more logical to pick an element whose rank is close to $\alpha\cdot s$ in the sample, if we have to select the element of rank $m=\alpha\cdot n$. We call this sampling strategy *proportional-of-s*. This is also the idea behind Floyd and Rivest's SELECT algorithm [2] whose expected cost is optimal, namely, $\lim_{n\to\infty,m/n\to\alpha} C_{n,m}/n = 1+\min(\alpha,1-\alpha)$; their algorithm uses two wisely chosen pivots from a variable-size sample at each recursive stage.

There are no previous results about the performance of quickselect with proportional-of-s sampling, nor even has this variant been formally proposed, to the best of the authors' knowledge. We tackle in this paper its analysis, beginning with the the recurrence for the average cost of proportional-of-s sampling and a characterization of $f(\alpha) = \lim_{n\to\infty,m/n\to\alpha} C_{n,m}/n$ as the fixed-point of an integral equation in Section 2. In Sections 3 and 4 we explicitly solve the equations for the proportional-of-2 and proportional-of-3 strategies. Our analysis shows that for proportional-of-3 $f(\alpha)\le m_1(\alpha)$ for all α with the exception of $0.201\ldots<\alpha<1/3$ and $2/3<\alpha<0.798\ldots$. In Section 5 we consider a variant of proportional-of-3 where we choose the smallest in the sample if $\alpha\le\nu$, the median if $\nu<\alpha<1-\nu$ and the largest if $\alpha\ge 1-\nu$, and explore various parameters as ν varies. We provide the optimal ν and a range of values of ν such that this variant is always better than median-of-3. Since the only changes to the basic algorithm involve the selection of the pivot for each recursive stage, these variants can be easily implemented. Finally, in Section 6 we briefly comment on further research.

2 General results

We begin this section with the derivation of the integral equation that satisfies $f(\alpha) = \lim_{n\to\infty, m/n\to\alpha} C_{n,m}/n$ when we use the proportional-of-s sampling. Actually, we generalize these results to a broader class of algorithms that we call *adaptive sampling strategies*.

We consider that, at each stage, the rank r of the selected pivot is a function of $\alpha = m/n$, the ratio of the current rank m to the current size n. Let $\pi_{n,j}^{(s,r)}$ be the probability that the rth element of a sample of s elements is the jth element of a random permutation of size n. Clearly,

$$\pi_{n,j}^{(s,r)} = \frac{\binom{j-1}{r-1}\binom{n-j}{s-r}}{\binom{n}{s}}, \qquad 1 \le r \le s \le n, \quad 1 \le j \le n. \tag{1}$$

The denominator is the number of ways to pick a sample of size s out of n elements; the numerator is the number of ways to choose $r-1$ elements smaller than the pivot times the number of ways to choose $s-r$ elements larger than the pivot.

Now we are ready to set up a recurrence for the average number of comparisons made to select the mth element out of n: since $n-1$ comparisons are needed to partition the array around the pivot [1] and we stop if the pivot is the sought element, we have

$$C_{n,m} = n - 1 + \sum_{j=m+1}^{n} \pi_{n,j}^{(s,r)} \cdot C_{j-1,m} + \sum_{j=1}^{m-1} \pi_{n,j}^{(s,r)} \cdot C_{n-j,m-j}. \tag{2}$$

We assume that $r(\alpha)$ is an integer staircase function defined by a finite collection of ℓ steps. We say that a sampling strategy defined by such a function is *adaptive*. The functions that do not satisfy the assumption would behave rather strangely and are most likely useless.

Hence r can be described by the image of each interval in a finite partition of $(0,1)$, say $I_1, I_2, \ldots, I_\ell$ with endpoints $a_0 = 0 < a_1 < a_2 < \cdots < a_\ell = 1$; we denote r_k the value of r for $\alpha \in I_k$. For convenience, we assume $I_k = (a_{k-1}, a_k]$ if $a_k \le 1/2$, $I_k = [a_{k-1}, a_k)$ if $a_{k-1} > 1/2$ and $I_k = (a_{k-1}, a_k)$ if $a_{k-1} \le 1/2 < a_k$.

For instance, median-of-$(2t+1)$ is defined by a single interval ($\ell = 1$) and $r_1 = t+1$: no matter what the value of α is, we always choose the median of the sample.

On the other hand, we can now formally define *proportional-of-s sampling*: it is the sampling strategy defined by s intervals, with $a_k = k/s$ and $r_k = k$. For

[1] We do not count the comparisons required to select the pivot since $s = \Theta(1)$.

instance, if $s = 3$ then $r(\alpha) = 1$ for $\alpha \in (0, 1/3]$, $r(\alpha) = 2$ for $\alpha \in (1/3, 2/3)$ and $r(\alpha) = 3$ if $\alpha \in [2/3, 1)$.

We state now the main result of this section.

Theorem 1. *Let $C_{n,m}$ be the average cost to select the mth out of n elements using an adaptive sampling strategy with $m/n \to \alpha$ for $0 < \alpha < 1$ as $n \to \infty$. Then we have*

$$f(\alpha) = \lim_{n\to\infty} \frac{C_{n,m}}{n}$$

with $f(\alpha) = f_k(\alpha)$ if $\alpha \in I_k$, $1 \le k \le \ell$ and

$$\begin{aligned} f_k(\alpha) = 1 + \frac{s!}{(r_k-1)!(s-r_k)!} \Bigg[&\int_{\alpha/a_k}^{1} f_k(\alpha/x)\, x^{r_k} (1-x)^{s-r_k}\, dx \\ &+ \int_0^{\frac{\alpha-a_{k-1}}{1-a_{k-1}}} f_k\left(\frac{\alpha-x}{1-x}\right) x^{r_k-1}(1-x)^{s+1-r_k}\, dx \\ &+ \sum_{d=k+1}^{\ell} \int_{I'_d} f_d(\alpha/x)\, x^{r_k}(1-x)^{s-r_k}\, dx \\ &+ \sum_{d=1}^{k-1} \int_{I''_d} f_d\left(\frac{\alpha-x}{1-x}\right) x^{r_k-1}(1-x)^{s+1-r_k}\, dx, \Bigg] \end{aligned}$$

with $I'_d = (\alpha/a_d, \alpha/a_{d-1})$ and $I''_d = \left(\frac{\alpha-a_d}{1-a_d}, \frac{\alpha-a_{d-1}}{1-a_{d-1}}\right)$.

Proof (Sketch). The proof of this theorem closely follows the proof in [3] for the median-of-$(2t+1)$ variant. The technical details of the proof are involved, but the intuition behind is rather simple: divide both sides of (2) by n, take $j = x \cdot n$ in the summations, and substitute $C_{j-1,m}$ by $f(\alpha/x) \cdot x \cdot n$ and $C_{n-j,m-j}$ by $f((\alpha-x)/(1-x)) \cdot (1-x) \cdot n$. Also, when $n \to \infty$ and $\alpha \in I_k$ we can replace $n \cdot \pi_{n,j}^{(s,r_k)}$ by the asymptotic estimate

$$n \cdot \pi_{n,j}^{(s,r_k)} \to \frac{s!}{(r_k-1)!(s-r_k)!} x^{r_k-1}(1-x)^{s-r_k} = \mathrm{Beta}(r_k, s+1-r_k).$$

This means, in the terminology of [3, 4] that the μ-split for a particular adaptive sampling strategy in its kth interval is built from the $\mathrm{Beta}(r_k, s+1-r_k)$-split. Passing to the limit when $n \to \infty$, the summations become integrals thus yielding the equations stated in the theorem. Since r is an adaptive function, the range of α must be broken into ℓ intervals, giving the piecewise definition of $f(\alpha)$. □

We say that an adaptive sampling strategy is *symmetric* if $\lim_{z\to\alpha^-} r(1-z) = s + 1 - \lim_{z\to\alpha^-} r(z)$ for all α. This definition properly captures the symmetric nature of the algorithm. Indeed, if we use as a pivot the rth smallest element in the sample of size s when searching for the element with rank α, it is reasonable to choose the rth largest (equivalently, the $(s+1-r)$th smallest) element while searching for the element of rank $1-\alpha$. The actual definition using limits from the left is necessary to make it valid also for samples of even size. Notice that for any symmetric sampling strategy we have $a_k = 1 - a_{\ell-k}$ if $k < s/2$; furthermore if the number of intervals ℓ is even then $a_{\ell/2} = 1/2$.

We immediately obtain the following lemmas (due to space limitations, we do not include proofs in this extended abstract).

Lemma 1. *For any symmetric adaptive sampling strategy, we have* $f(\alpha) = f(1-\alpha)$. *More precisely,* $f_k(\alpha) = f_{\ell+1-k}(1-\alpha)$, *if* $\alpha \in I_k$.

Lemma 2. *Let* $r_0 = \lim_{\alpha\to 0^+} r(\alpha)$. *For any adaptive sampling strategy we have*

$$\lim_{\alpha\to 0^+} f(\alpha) = \frac{s+1}{s+1-r_0}.$$

Now we can easily rederive the known fact [3] that for median-of-$(2t+1)$, $\lim_{\alpha\to 0^+} m_t(\alpha) = ((2t+1)+1)/((2t+1)+1-(t+1)) = 2$, for any t. On the other hand, if we use proportional-of-s sampling or a similarly inspired strategy such that $r_0 = 1$, then $\lim_{\alpha\to 0^+} f(\alpha) = 1 + 1/s$, which proves that significant gains can be expected from alternative sampling strategies, even for $s = 2$.

In order to provide an explicit solution of the integral equations in Theorem 1, we transform the original problem to one of ordinary differential equations.

Lemma 3. *For* $\alpha \in I_k$, $1 \le k \le \ell$, *and any adaptive sampling strategy*

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

Since proportional-of-s is a symmetric strategy, we only have to consider the equations for $1 \le k \le \lceil s/2 \rceil$ and the order of the ordinary differential equation satisfied by each f_k can be reduced. Let $\phi_k(x) = d^{k+1} f_k / dx^{k+1}$. Then, for all $1 \le k \le \lceil s/2 \rceil$

$$\frac{d^{s+1-k}}{dx^{s+1-k}}\phi_k - \frac{s!}{(s-k)!}\frac{1}{(1-x)^k}\frac{d^{s+1-2k}}{dx^{s+1-2k}}\phi_k - \frac{s!}{(k-1)!}\frac{(-1)^{s+1-k}}{x^{s+1-k}}\phi_k(x) = 0. \quad (3)$$

An important special case of the ODE above is for the central interval ($k = (s+1)/2$) when s is odd. The ODE is then identical to the corresponding ODE for median-of-s.

The problem with the differential equations above, besides the intrinsic difficulty of solving high order linear differential equations, is that initial conditions are hard to establish, other than the limiting value $f(0)$. Recall that $f(\alpha)$ is in general discontinuous and hence in order to obtain it, we should know $f_k(a_{k-1})$, $f'_k(a_{k-1})$, …, $f_k^{(s+1)}(a_{k-1})$ for every k, $1 \le k \le \ell$. In order to overcome the problem, we use a different technique: namely, substitute the f_k's in the integral equation(s) by the general form of the solution to the corresponding differential equation and fix the values of the involved constants by equalizing both sides. When the strategy is symmetric the problem is somewhat simpler, but the essential obstacles remain.

3 Proportional-of-2

Let us begin with the simplest "proportional-of" strategy: $s = 2$. Solving (3) is not very difficult in this case and even can be said to be routine. Here, we have just to consider one piece, namely ϕ_1, since by symmetry we know that $f_2(x) = f_1(1 - x)$. Equation (3) is then

$$\frac{d^2\phi_1}{dx^2} - \frac{2}{1-x}\frac{d\phi_1}{dx} - \frac{2}{x^2}\phi_1(x) = 0, \tag{4}$$

with $x = 0$ and $x = 1$ its regular singular points. We remark that $\phi_1(x) = d^2 f_1/dx^2$. The corresponding indicial equation is $\lambda(\lambda-1)-2 = 0$ whose solutions are $\lambda = 2$ and $\lambda = -1$. This entails a solution of the form $\phi_1(x) = y_1(x) + y_2(x)$ where

$$y_1(x) = \sum_{n\geq 0} A_n x^{n+2}, \qquad y_2(x) = \sum_{n\geq 0} B_n x^{n-1} + A \cdot y_1(x) \cdot \ln x,$$

for some coefficients $\{A_n\}_{n\geq 0}$ and $\{B_n\}_{n\geq 0}$. Substituting the proposed form for $\phi_1(x)$ into the differential equation and integrating ϕ_1 twice, we finally get

$$f_1(x) = a((x-1)\ln(1-x) + \frac{1}{6}x^3 + \frac{1}{2}x^2 - x) - b(1 + \mathcal{H}(x)) + cx + d,$$

for some constants a, b, c and d yet to be determined.

The difficult part here is to obtain the values of these constants. The known value $f_1(0) = 3/2$ gives $d - b = 3/2$, but the successive derivatives of f_1 at $\alpha = 0$ are infinite and this information cannot be used to fix the value of the remaining constants. The painful process is to substitute the general expression for f_1 and f_2 into the integral equation (Theorem 1) and equalize.

Finally, one gets

$$a = -2 + \frac{3}{2}\frac{4\ln 2 - 3}{3\ln 2 - 2}, \qquad b = -2, \qquad c = -\frac{1}{8}\frac{4\ln 2 - 3}{3\ln 2 - 2}, \qquad d = -1/2.$$

The maximum of $f(\alpha)$ is at $\alpha = 1/2$. Indeed, we have $f'(0) = \infty$ and $f'(1/2) > 0$. It is easy to check, differentiating one more time, that $f'(\alpha)$ is strictly decreasing, and hence always positive. In $\alpha = 1/2$, the cost is $f_1(1/2) = f_2(1/2) = 1/96 \cdot (576 \ln^2 2 - 253)/(3\ln 2 - 2) \doteq 3.112\ldots$. Compare with $m_0(1/2) \doteq 3.386\ldots$ for standard quickselect and $m_1(1/2) = 11/4 = 2.75$ for median-of-3 quickselect[2].

[2] We only give four figures in the numerical values in this section and the following. But it is relatively easy to obtain a high degree of accuracy—indeed, we have computed all the numerical values given in the paper with up to twenty digits of accuracy.

Also, we have

$$\overline{f} = \int_0^1 f(x)\,dx = 2 \cdot \int_0^{1/2} f_1(x)\,dx = \frac{3}{128}\,\frac{320\ln 2 - 213}{3\ln 2 - 2} \doteq 2.598\ldots,$$

which tells us that proportional-of-2 makes $2.6n$ comparisons on average. Compare with the $3n$ comparisons of the standard variant and the $2.5n$ comparisons of the median-of-3 variant.

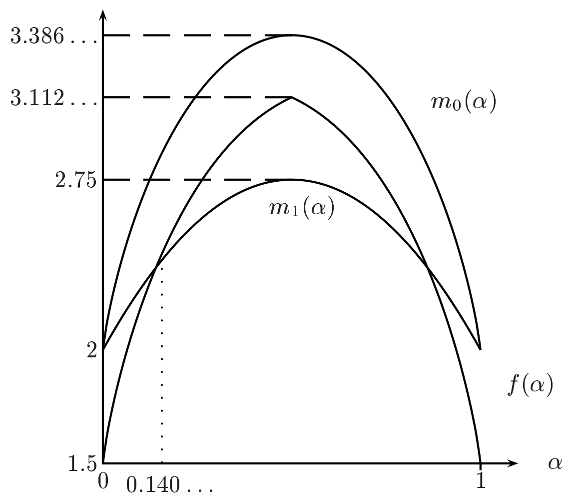

Fig. 1. Plot of $m_0(\alpha)$, $m_1(\alpha)$ and $f(\alpha)$ for proportional-of-2

It is interesting to notice that $f(\alpha) \leq m_0(\alpha)$ for all $0 \leq \alpha \leq 1$. Compared to median-of-3, proportional-of-2 is better when $\alpha \leq 0.140\ldots$ (symmetrically, when $\alpha \geq 0.859\ldots$ and is worse otherwise (see Figure 1). In that percentile, both algorithms make in average $2.362\ldots n$ comparisons. However, it is a bit unfair to compare median-of-3 and proportional-of-2 since these strategies use a different number of elements in the samples (and standard quickselect uses samples of size $s = 1$!).

4 Batfind: Proportional-of-3

The steps needed to analyze *batfind* (a nickname for proportional-of-3) are similar to those of the previous section. In this case we have the following three functions: $f_1(x)$ when $x \in (0, 1/3]$, $f_2(x)$ when $x \in (1/3, 2/3)$ and $f_3(x)$ when $x \in [2/3, 1)$. By symmetry we have $f_3(x) = f_1(1-x)$ and $f_2(x) = f_2(1-x)$.

This implies that we need to solve only two differential equations, namely,

$$\frac{d^3\phi_1}{dx^3} - \frac{3}{1-x}\frac{d^2\phi_1}{dx^2} - \frac{6}{x^3}\phi_1(x) = 0, \qquad \frac{d^2\phi_2}{dx^2} - 6\left(\frac{1}{x^2} + \frac{1}{(1-x)^2}\right)\phi_2(x) = 0, \tag{5}$$

with $\phi_1(x) = d^2 f_1/dx^2$, $\phi_2(x) = d^3 f_2/dx^3$, and $x = 0$ and $x = 1$ their regular singular points. Using the same procedure as in Section 3 we get

$$f_1(x) = -C_0(1 + \mathcal{H}(x)) + C_1 + C_2 x + C_3 \cdot K_1(x) + C_4 \cdot K_2(x)$$

where

$$K_1(x) = \cos\left(\sqrt{2}\ln x\right) \cdot \sum_{n\geq 0} A_n\, x^{n+4} + \sin\left(\sqrt{2}\ln x\right) \cdot \sum_{n\geq 0} B_n\, x^{n+4},$$

$$K_2(x) = \sin\left(\sqrt{2}\ln x\right) \cdot \sum_{n\geq 0} A_n\, x^{n+4} - \cos\left(\sqrt{2}\ln x\right) \cdot \sum_{n\geq 0} B_n\, x^{n+4},$$

$$A_n = \frac{(n+2)(n+5)}{(n^2+6n+11)(n^2+8n+18)}, \qquad B_n = \frac{\sqrt{2}(2n+7)}{(n^2+6n+11)(n^2+8n+18)}.$$

We also obtain

$$f_2(x) = -C_5(1 + \mathcal{H}(x)) + C_6 x(1-x) + C_7.$$

The value of the constants can be obtained by the same routine but cumbersome procedure of substitution into the integral equations. We get thus

$$C_0 = -24/11, C_1 = -28/33, C_2 \doteq 0.193\ldots, C_3 \doteq -100.190\ldots,$$

$$C_4 \doteq -27.556\ldots, C_5 \doteq -1.463\ldots, C_6 = \frac{7}{4}C_5 + 3 \doteq 0.439\ldots, C_7 \doteq 0.135\ldots.$$

We observe that, contrary to what happens with $m_1(x)$, $f_2(x)$ is the sum of a second degree polynomial and an entropic term (recall that the linear differential equation is the same in both cases, but the entropic term vanishes for median-of-3). Another important aspect is that, even though the difference is not large, $f(\alpha) \leq m_1(\alpha)$ for $1/3 \leq \alpha \leq 2/3$, and the same is true for the intervals $\alpha \leq 0.201\ldots$ and $\alpha \geq 0.798\ldots$ In particular, $f(1/2) = f_2(1/2) \doteq 2.722\ldots < m_1(1/2) = 2.75$.

Finally, integrating $f(\alpha)$ in $(0,1)$ yields $\overline{f} \doteq 2.421\ldots$ which favorably compares to the value $\overline{m}_1 = 2.5$ that corresponds to median-of-3.

We conclude this section briefly discussing the intuition behind the fact that batfind is doing worse than median-of-3 for values of α near $1/3$ (and $2/3$). It also makes more comparisons in these regions than for $\alpha = 1/2$. In particular, if $\alpha \in [0.276\ldots, 1/3]$ or $\alpha \in [2/3, 0.723\ldots]$ then batfind makes more comparisons, on the average, to select the element of rank $\alpha \cdot n$ than to select the median.

An informal explanation is the following. Assume for the sake of concreteness that $s = 3$, $n = 1000$ and $m = 332$. While there is some chance that the rank of the pivot selected by batfind is close but larger than $1000/3 = 333.3$ —then

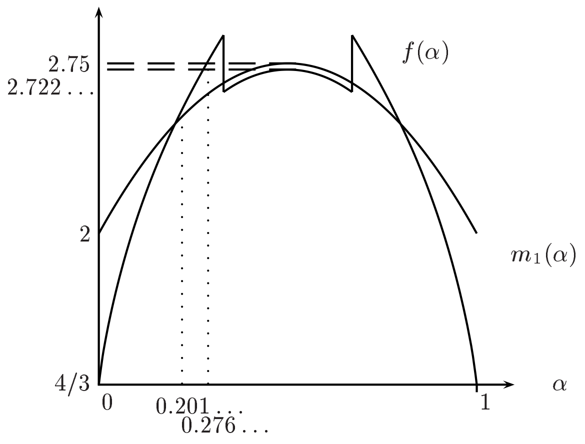

Fig. 2. Plot of batfind's $f(\alpha)$ and $m_1(\alpha)$.

we are in good shape since we have discarded almost two thirds of the input—, it is also likely that the rank of the pivot is slightly less than m and then we have discarded a bit less than a third of the input. In the next round the rank of the sought element will be relatively small; however there are still enough chances that we have "bad luck" again, as in the first round. On the other hand, if $m = 334$ then the strategy would pick the median of the sample and thus exhibit more "stable" performance, since it would most likely partition the array into subarrays of similar size and hence avoid the boundary effect just described. Such boundary effects occurring at early stages of the recursion have a big impact on the performance and amount for the difficulty of finding elements whose rank is slightly less than $n/3$.

5 ν-find: A variant of batfind

In the following we consider symmetric sampling strategies that are inspired by proportional-of-s but try to overcome its deficiencies. For these strategies we have in general $a_k \neq k/s$ but $r_k = k$ for all $1 \leq k \leq s$. One important point is that these variants of proportional-of-s satisfy the same differential equations. Hence, the general form of the f_k's is exactly the same and the only difference is in the value of the involved constants, because of the different initial conditions.

Here, we only explore the case where $s = 3$ and investigate the properties of $f(\alpha)$ when $a_1 = \nu$ and $a_2 = 1 - \nu$ for $0 < \nu < 1/2$; we make the dependence in ν explicit and denote f_ν the function corresponding to this strategy, which we call ν-find. For ν-find the differential equations are the same as in (5). When $\nu \to 0$, ν-find behaves as quickselect with median-of-three. When $\nu = 1/3$, we recover batfind. Finally, when $\nu \to 1/2$, the median of the sample is always discarded, so ν-find behaves slightly different than proportional-of-2.

In general, $f_\nu(\alpha)$ consists of three pieces: $f_{1,\nu}$ for $\alpha \in (0,\nu]$, $f_{2,\nu}$ for $\alpha \in (\nu, 1-\nu)$ and $f_{3,\nu}$ for $\alpha \in [1-\nu, 1)$. Of course, since ν-find is symmetric we have $f_{3,\nu}(\alpha) = f_{1,\nu}(1-\alpha)$.

As we have already pointed out, the only difference between our analysis of batfind and that of ν-find is that we have to obtain the dependence on ν of the values of the constants C_i in the general form of f_1 and f_2 (see Section 4). Notice that the argument of symmetry of $f_{2,\nu}$ applies also here, no matter what the value of ν is.

It turns out that $C_0 = -24/11$ and $C_1 = -28/33$ are independent of ν. Moreover, $C_6(\nu) = 7/4 \cdot C_5(\nu) + 3$. The formulæ for the other C_i's are complicated (four to five lines each) and we do not give them here. Also, due to the space limitations, the rest of this section will describe the most salient features of ν-find, but we are not able to give the corresponding proofs or calculations here.

For a large range of values of ν, f_ν has three local maxima located at $\alpha = \nu$, $\alpha = 1-\nu$ and $\alpha = 1/2$. The local maxima at ν and $1-\nu$ constitute the so characteristic two little "ears" of ν-find (and batfind in particular).

As ν decreases, the value of $f_{1,\nu}(\nu)$ also decreases and $\alpha = 1/2$ becomes the absolute maximum of f_ν. We denote $\tilde{\nu}$ the largest value of ν such that $\alpha = 1/2$ is the absolute maximum of f_ν; for $\nu > \tilde{\nu}$ the absolute maxima of f_ν are located at $\alpha = \nu$ and $\alpha = 1-\nu$. So $\tilde{\nu}$ is the solution of $f_\nu(\nu) = f_\nu(1/2)$. Numerical computations yield $\tilde{\nu} \doteq 0.268\ldots$.

If we continue decreasing the value of ν then f_ν loses its characteristic "ears" because then we have $f_{1,\nu}(\nu) < f_{2,\nu}(\nu)$ (see Figure 3). We denote $\hat{\nu}$ the transition point where $f_{1,\nu}(\nu) = f_{2,\nu}(\nu)$, i.e., where f_ν is continuous. We have that $\hat{\nu} \doteq 0.182\ldots$.

This transition point enjoys another fundamental property: it is the optimal value $\hat{\nu} = \nu^*$ since $f_{\nu^*}(\alpha) \leq f_\nu(\alpha)$, for any ν and any α, and no other value of ν has this property.

Since ν^* optimizes f_ν, it minimizes the maxima; furthermore $\nu^* < \tilde{\nu}$, so it follows that ν^* minimizes $f_\nu(1/2)$. We can characterize it by $\partial f_\nu(1/2)/\partial\nu|_{\nu=\nu^*} = 0$. In particular, $f_{\nu^*}(1/2) \doteq 2.659\ldots$ (see Figure 4).

It is also obvious that ν^* must minimize the average value $\overline{f}_\nu$. Therefore, we have that $\partial\overline{f}_\nu/\partial\nu|_{\nu=\nu^*} = 0$. The minimum value is $\overline{f}_{\nu^*} \doteq 2.342\ldots$.

Another set of interesting values of ν appear when we compare ν-find with median-of-3. In particular:

1. For $\nu \leq \overline{\nu}' \doteq 0.404\ldots$, ν-find does better than median-of-3 on the average; that is, $\overline{f}_\nu \leq \overline{m}_1 = 5/2 = 2.5$.
2. For $\nu \leq \nu'_m \doteq 0.364\ldots$, ν-find does better than median-of-3 to locate the median, or in other words, $f_\nu(1/2) \leq m_1(1/2) = 11/4 = 2.75$ (see Figure 4).
3. For $\nu \leq \nu' \doteq 0.219\ldots$, ν-find does always better than median-of-3, that is $f_\nu(\alpha) \leq m_1(\alpha)$ for all α. Because of the properties of f_ν and m_1, ν' is characterized as the solution of $m_1(\nu) - f_{1,\nu}(\nu) = 0$. Notice that $\nu' \leq \tilde{\nu}$ so that when ν-find beats median-of-3, $\alpha = 1/2$ is already the most difficult rank to be selected for ν-find; but on the other hand, $f_{1,\nu'}(\nu') > f_{2,\nu'}(\nu')$ (see Figure 4).

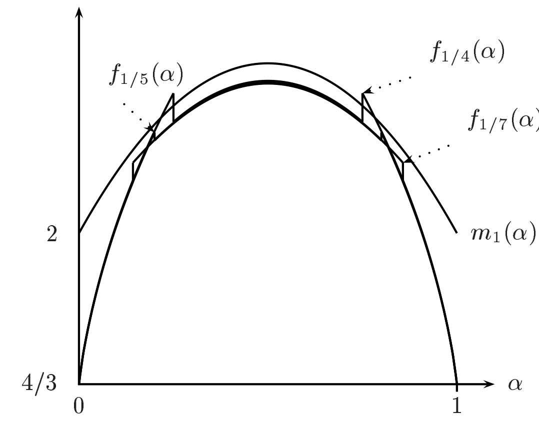

Fig. 3. Plot of $f_\nu(\alpha)$ for $\nu \in \{1/4, 1/5, 1/7\}$. Median-of-three ($m_1(\alpha)$) is also depicted for comparison.

4. For $\nu > \nu'$, by definition, ν-find does worse than median-of-three for some intervals of α. In particular, if $\nu \leq \nu'_m$ then ν-find beats median-of-3 in $(0, \alpha_\nu]$, $(\nu, 1-\nu)$ and $[1-\alpha_\nu, 1)$; if $\nu > \nu'_m$ then ν-find beats median-of-3 only in $(0, \alpha_\nu]$ and $[1-\alpha_\nu, 1)$. For instance, since $\nu' < 1/3 < \nu'_m$, batfind beats median-of-3 in the ranges $(0, \alpha_{1/3}]$, $(1/3, 2/3)$ and $[1-\alpha_{1/3}, 1)$, with $\alpha_{1/3} \doteq 0.201\ldots$. In general, α_ν is the solution of $m_1(\alpha_\nu) - f_{1,\nu}(\alpha_\nu) = 0$.

6 Future work

Although it seems that it would not be too difficult to extend our analysis of proportional-of-s for $s > 3$, say $s = 4$ and $s = 5$, the process is undoubtely tedious and the results most likely cumbersome and qualitatively similar. A more interesting open problem is to provide the analysis of proportional-of-s as $s \to \infty$.

On the other hand, we are considering randomized sampling strategies, where given the relative rank α of the sought element, for each r, $1 \leq r \leq s$, there is a probability $p_r(\alpha)$ that the rth smallest element of the sample is chosen as the pivot. These strategies generalize the deterministic strategies studied in this paper.

References

1. D.H. Anderson and R. Brown. Combinatorial aspects of C.A.R. Hoare's FIND algorithm. *Australasian Journal of Combinatorics*, 5:109–119, 1992.
2. R.W. Floyd and R.L. Rivest. Expected time bounds for selection. *Communications of the ACM*, 18:165–173, 1975.

2.82
2.72
2.62
2.52
2.42
2.32
2.22
2.12
2.2
0.15 0.20 0.25 0.30 0.35

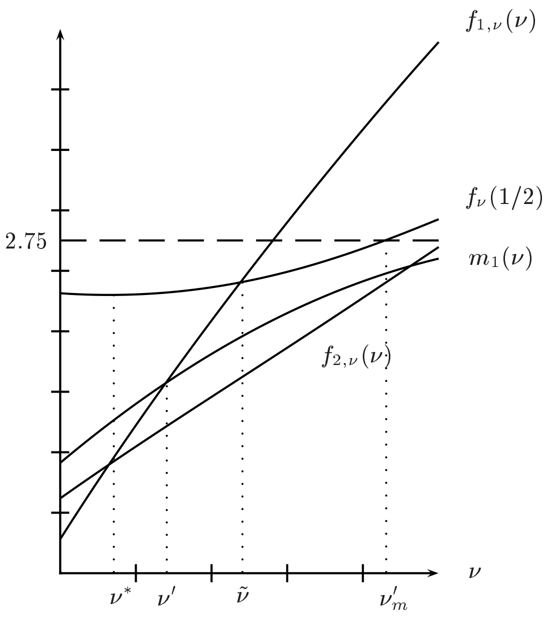

Fig. 4. Plot of $f_{1,\nu}(\nu)$, $f_{2,\nu}(\nu)$, $f_\nu(1/2)$ and $m_1(\nu)$, and several interesting values of ν.

3. R. Grübel. On the median-of-k version of Hoare's selection algorithm. *Theoretical Informatics and Applications*, 33(2):177–192, 1999.
4. R. Grübel and U. Rösler. Asymptotic distribution theory for Hoare's selection algorithm. *Adv. Appl. Prob.*, 28:252–269, 1996.
5. C.A.R. Hoare. FIND (Algorithm 65). *Communications of the ACM*, 4:321–322, 1961.
6. C.A.R. Hoare. Quicksort. *Computer Journal*, 5:10–15, 1962.
7. P. Kirschenhofer, H. Prodinger, and C. Martínez. Analysis of Hoare's FIND algorithm with median-of-three partition. *Random Structures & Algorithms*, 10:143–156, 1997.
8. D.E. Knuth. Mathematical analysis of algorithms. In *Information Processing '71, Proc. of the 1971 IFIP Congress*, pages 19–27, Amsterdam, 1972. North-Holland.
9. D.E. Knuth. *The Art of Computer Programming: Sorting and Searching*, volume 3. Addison-Wesley, Reading, Mass., 2nd edition, 1998.
10. H.M. Mahmoud. *Sorting: A Distribution Theory.* John Wiley & Sons, New York, 2000.
11. C. Martínez and S. Roura. Optimal sampling strategies in quicksort and quickselect. *SIAM Journal on Computing*, 31(3):683–705, 2001.