PEDECIBA Informática
Instituto de Computación – Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay

Reporte Técnico RT 04-14

Adaptative sampling strategies for Quickselect

Conrado Martínez   Daniel Panario   Alfredo Viola

Julio de 2004

Adaptative sampling strategies for Quickselect
Martínez, Conrado; Panario, Daniel; Viola, Alfredo.

ISSN 0797-6410
Reporte Técnico RT 04-14
PEDECIBA
Instituto de Computación – Facultad de Ingeniería
Universidad de la República

Montevideo, Uruguay, julio de 2004

Adaptive Sampling Strategies for Quickselect[*]

Conrado Martínez[†] Daniel Panario[‡] Alfredo Viola[§]

July 17, 2004

Abstract

Quickselect with median-of-3 is largely used in practice and its behavior is fairly well understood. However, the following natural adaptive variant, which we call *proportion-from-3*, had not been previously analyzed: "choose as pivot the smallest of the sample if the relative rank of the sought element is below 1/3, the largest if the relative rank is above 2/3, and the median if the relative rank is between 1/3 and 2/3". We first analyze the average number of comparisons made when using proportion-from-2 and then for proportion-from-3. We also analyze ν-find, a generalization of proportion-from-3 with interval breakpoints at ν and $1-\nu$. We show that there exists an optimal value of ν and we also provide the range of values of ν where ν-find outperforms median-of-3. Then, we consider the average total cost of these strategies, which takes into account the cost of both comparisons and exchanges. Our results strongly suggest that a suitable implementation of ν-find could be the method of choice in a practical setting. We also study the behavior of proportion-from-s with $s > 3$ and in particular we show that proportion-from-s-like strategies are optimal when $s \to \infty$.

1 Introduction

Hoare's quickselect [9] finds 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

[*] 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 the Spanish Min. of Science and Technology project TIC2002-00190 (AEDRI II). The second author was funded by NSERC of Canada grant number 238757. The third author was supported in part by Proyecto de Investigación CSIC fondos 2000–2002 and 2002–2004. Part of the research of the third author was done while he held a CNRS position at LIPN, Université de Paris-Nord, France.

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

[‡] School of Mathematics and Statistics. Carleton University. K1S 5B6, Ottawa, Canada. `daniel@math.carleton.ca`.

[§] Instituto de Computación. Universidad de la República. Casilla de Correo 16120, Distrito 6, Montevideo, Uruguay. `viola@fing.edu.uy`.

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 (see Algorithm 1). A similar principle is used in the celebrated quicksort algorithm [10], also by Hoare; once the pivot is brought into place by partitioning the array, the subarrays to its left and right are recursively sorted.

Algorithm 1 The quickselect algorithm.

```
{$l \leq m \leq u$, $A[l..u]$ contains the $m$th smallest of $A[1..n]$ }

function quickselect(var $A$ : array $[1..n]$ of Elem; $m, l, u$ : integer)
    var $r, j$ : integer;
    begin
        if $l = u$ then return $A[l]$;
        $r :=$ pick_pivot($A, l, u, m$);
        swap($A[l], A[r]$);

        {$A[l] = p$}
        partition($A, l, u, j$);
        {$\forall i : l \leq i < j : A[i] \leq p$, $A[j] = p$, and $\forall i : j < i \leq u : A[i] > p$}

        if $m = j$ then return $A[j]$;
                else if $m < j$ then return quickselect($A, m, l, j-1$);
                              else return quickselect($A, m, j+1, u$)
    end
```

For the remaining of this paper we use $\alpha = m/n$ to denote the *relative* rank of the sought element. Quickselect performs well in practice, its average cost being linear. In particular, if $C^{(0)}_{n,m}$ denotes the average number of comparisons made by quickselect to select the mth element out of n, and $\mathcal{H}(x) = -x \ln x - (1-x)\ln(1-x)$ is the entropy function, we have (see [8, 13, 14, 15])

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

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 $C^{(0)}_n = 1/n \cdot \sum_{1\leq m\leq n} C^{(0)}_{n,m}$, we have [15]

$$\overline{m}_0 = \lim_{n\to\infty} \frac{C^{(0)}_n}{n} = 3.$$

Using the median of a small sample of $2t+1$ 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 [18]. For samples of three elements [1, 7, 12] 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 quickselect with median-of-$(2t+1)$ sampling to select the mth out of n elements. Also,

$$\overline{m}_1 = \lim_{n\to\infty} \frac{C^{(1)}_{n}}{n} = \frac{5}{2}.$$

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. In general, it seems better and more logical to pick an element whose rank is close to $\alpha \cdot s$ out of a sample of size s, if we have to select the element of rank $m = \alpha \cdot n$. We call this sampling strategy *proportion-from-s*. A similar idea lies behind Floyd and Rivest's SELECT algorithm [5]; this algorithm sorts a variable-size sample at each recursive stage to obtain two pivots. Because of the costly selection of pivots, the algorithm is not very efficient in practice, although it is within a lower order term of the theoretical optimal (see also [4]). Its expected cost satisfies $\lim_{n\to\infty, m/n\to\alpha} C_{n,m}/n = 1 + \min(\alpha, 1-\alpha)$. But despite the optimal performance of SELECT and some other similar algorithms, in practice finely tuned implementations of Hoare's quickselect with median-of-3 sampling are preferred and used in system and general-purpose libraries such as the Standard Template Library of C++ (see for instance [19]).

There are no previous results about the performance of quickselect with proportion-from-s sampling, nor even had this variant been formally proposed, to the best of the authors' knowledge. We tackle in this paper its analysis, beginning in Section 2 with the recurrence for the average cost of proportion-from-s sampling and the integral equation satisfied by the *characteristic function* $f(\alpha) = \lim_{n\to\infty, m/n\to\alpha} C_{n,m}/n$ of the proportion-from-s algorithm. In Sections 3 and 4 we explicitly solve the equations for the proportion-from-2 and proportion-from-3 strategies and investigate some of their properties. We also discuss briefly the general form of the solution for general s. In Section 5 we consider a variant of proportion-from-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. Afterwards, in Section 6 we consider the average number of exchanges and the average total cost of ν-find. Since our analysis does only consider the main order term in the cost of quickselect in the asymptotic regime, we have conducted a few experiments to assess the validity of the theoretical predictions in practical terms; these experiments and their results are described in Section 7. In Section 8 we prove that proportion-from-s-like sampling strategies achieve optimal performance when $s \to \infty$. We conclude in Section 9 with future research directions and open problems.

A preliminary version of this paper has appeared in [17].

2 General results

We begin this section with the derivation of the integral equation satisfied by $f^{(s)}(\alpha) = \lim_{n\to\infty, m/n\to\alpha} C_{n,m}/n$ when we use proportion-from-s sampling[1]. 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. Recall that we also use α to denote the initial relative rank, should no confusion arise.

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. First, $n-1$ comparisons are needed to partition the array around the pivot, and $\Theta(s) = \Theta(1)$ additional comparisons are necessary to select the pivot. If the pivot is the jth smallest element and $m < j$, we continue in the left subarray, still looking for the mth smallest element, but now the array contains $j-1$ elements. If $m > j$ then we continue in the right subarray which contains $n-j$ elements, but we look now for the $(m-j)$th smallest element there. Finally, we stop if the pivot is the sought element, i.e., when $j = m$. Hence we have

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

We assume that $r = r(\alpha)$ 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 set of disjoint intervals, say $I_1, I_2, \ldots, I_\ell$ with endpoints $a_0 = 0 < a_1 < a_2 < \cdots < a_\ell = 1$; we denote by r_k the value of r for $\alpha \in I_k$. For convenience, we assume $I_1 = [0, a_1]$, $I_\ell = [a_{\ell-1}, 1]$, $I_k = (a_{k-1}, a_k]$ if $k > 1$ and $a_k \le 1/2$, $I_k = [a_{k-1}, a_k)$ if $k < \ell$ and $a_{k-1} > 1/2$, and $I_k = (a_{k-1}, a_k)$ if $a_{k-1} \le 1/2 < a_k$ and $1 < k < \ell$. However, the forthcoming analysis is easily generalized for intervals defined in some different way, as long as the set of intervals totally covers $[0,1]$.

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 *proportion-from-s sampling*: it is the sampling strategy defined by s intervals, with $a_k = k/s$ and $r_k = k$. For instance, if

[1] If no confusion arises, for the rest of the paper we will drop the superscript denoting the sample size in $f^{(s)}(\alpha)$.

Algorithm 2 Picking a pivot with proportion-from-3 sampling.

```
function pick_pivot($(A, l, u, m)$)
    var $pl, pm, pu$ : integer; $n, rank$ : integer
    begin
        $n := u - l + 1$;
        if $n < 3$ then return $l$;
        {Any three distinct values in $[l..u]$ can be used to initialize $pl, pm, pu$}
        $pl := l$; $pm := (l + u)$ div $2$; $pu := u$;
        if $A[pl] > A[pm]$ then swap($pl, pm$);
        if $A[pm] > A[pu]$ then begin
                                  swap($pm, pu$);
                                  if $A[pl] > A[pm]$ then swap($pl, pm$);
                              end
        $rank := m - l + 1$;
        if $3 * rank \leq n$ then return $pl$;
        if $3 * rank < 2 * n$ then return $pm$;
        return $pu$;
    end
```

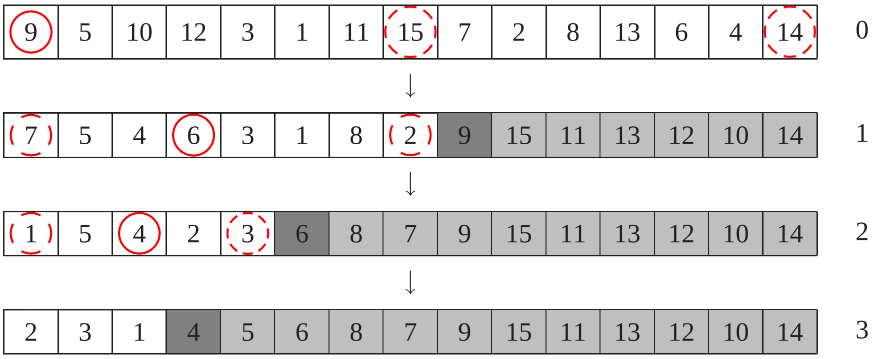
$s = 3$ then $r(\alpha) = 1$ for $\alpha \in [0, 1/3]$, $r(\alpha) = 2$ for $\alpha \in (1/3, 2/3)$ and $r(\alpha) = 3$ if $\alpha \in [2/3, 1]$. An example of an execution is presented in Figure 1, showing some interesting features of the algorithm. The most important observation is the fact that adaptive sampling strategies choose the pivot as a function of the current relative rank of the sought element at each recursive call. In this example, in the first recursive call we have to choose the first element ⑨ of the sample as the pivot because $\alpha = 4/15 < 1/3$, in the second recursive call the median of the sample ⑥ is selected since $1/3 < \alpha = 1/2 < 2/3$ and, finally, in the last recursive call we choose the largest element ④, as we have now $\alpha = 4/5 > 2/3$. In the figure, the pivot of each partition stage appears in dark grey once the partition is finished; discarded elements are shaded in light grey. The elements of the sample at each stage appear within a circle, which is solid if the sample element is the selected pivot.

The "reactiveness" of proportion-from-3 to the current relative rank of the sought element is the reason for its improved performance if compared with other variants of quickselect; but this simple modification of the pivot selection scheme makes the analysis of the average performance considerably difficult. However, the implementation of proportion-from-3 is as simple as for other variants of quickselect (see Algorithm 2). We observe that the selection of the pivot does not rearrange the elements of the sample in the array, but rather works with pointers (the indices pl, pm and pu) to them, for otherwise we would not preserve the randomness of the subarrays after partitioning.

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 $0 \leq \alpha \leq 1$ as $n \to \infty$. Then we have that the characteristic function of the algorithm*

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



9	5	10	12	3	1	11	15	7	2	8	13	6	4	14	0
7	5	4	6	3	1	8	2	9	15	11	13	12	10	14	1
1	5	4	2	3	6	8	7	9	15	11	13	12	10	14	2
2	3	1	4	5	6	8	7	9	15	11	13	12	10	14	3

Figure 1: An example of execution of the proportion-from-3 algorithm with $n = 15$ and $m = 4$.

with $f(\alpha) = f_k(\alpha)$ *if* $\alpha \in I_k$, $1 \le k \le \ell$ *is well defined, 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* $I''_d = \left(\frac{\alpha - a_d}{1-a_d}, \frac{\alpha - a_{d-1}}{1-a_{d-1}}\right)$.

Sketch of the proof. The proof is a generalization of the proofs in [7, 8] for standard quickselect and its median-of-$(2t+1)$ variants. First, we have to show that for any given partition $\{I_k\}$ of $[0,1]$ into ℓ disjoint intervals with endpoints $0 = a_0 < a_1 < \cdots < a_\ell = 1$ and any family $\{\mu_k\}_{1 \le k \le \ell}$ of probability distributions in $[0,1]$ which are not concentrated in $\{a_0, a_1, \ldots, a_\ell\}$ there is a unique family of probability distributions

$\{F_\alpha : 0 \leq \alpha \leq 1\}$ satisfying the following distributional equation[2]

$$F_\alpha \overset{\mathcal{L}}{=} 1 + 1_{(\alpha/a_k,1]}(\xi_k) \cdot \xi_k \cdot G_{\alpha/\xi_k} + 1_{\left[0,\frac{\alpha-a_{k-1}}{1-a_{k-1}}\right]}(\xi_k) \cdot (1-\xi_k) \cdot G_{(\alpha-\xi_k)/(1-\xi_k)}$$
$$+ \sum_{d=k+1}^{\ell} 1_{I'_d}(\xi_d) \cdot \xi_d \cdot G_{\alpha/\xi_d} + \sum_{d=1}^{k-1} 1_{I''_d}(\xi_d) \cdot (1-\xi_d) \cdot G_{(\alpha-\xi_d)/(1-\xi_d)},$$

where $\{G_\alpha : 0 \leq \alpha \leq 1\}$ is a family of independent random variables with the same distribution as $\{F_\alpha : 0 \leq \alpha \leq 1\}$ (i.e., $G_\alpha \overset{\mathcal{L}}{=} F_\alpha$, for all α), $\{\xi_k\}_{1\leq k\leq \ell}$ is a finite collection of independent random variables with the same distribution as the μ_k's, $I'_d = (\alpha/a_d, \alpha/a_{d-1})$, $I''_d = \left(\frac{\alpha-a_d}{1-a_d}, \frac{\alpha-a_{d-1}}{1-a_{d-1}}\right)$, and 1_I denotes the indicator function for the interval I. Thus we extend the notion of μ-split given in [7] to $\mu = \langle\{I_k\},\{\mu_k\}\rangle$ since such a pair uniquely determines the family of probability distributions $\{F_\alpha : 0 \leq \alpha \leq 1\}$. The result in [7] can be seen as the particular instance where $\ell = 1$ and thus $I_1 = [0,1]$.

The second part of the proof amounts to showing that for a given adaptive sampling strategy, if $\lim_{n\to\infty} m/n = \alpha$ then

$$\frac{C_{n,m}}{n} \xrightarrow{\mathcal{L}} F_\alpha$$

with respect to a suitable metric (in particular, with respect to the Wasserstein metric), where $C_{n,m}$ is the number of comparisons made by quickselect when selecting the mth smallest element out of n using the given adaptive sampling strategy, and $\{F_\alpha : 0 \leq \alpha \leq 1\}$ is the μ-split constructed with $\mu_k = \mathrm{Beta}(r_k, s+1-r_k)$ and the intervals corresponding to the given adaptive sampling strategy.

The statement of the theorem then easily follows. Also, as a by-product of the detailed proof, we can show that each $f_k(\alpha)$ is bounded in I_k, and furthermore establish that, for any s, the operator T_k defined by

$$T_k(g)(\alpha) := 1 + \frac{s!}{(r_k-1)!(s-r_k)!}\left[\int_{\alpha/a_k}^{1} g_k(\alpha/x)x^{r_k}(1-x)^{s-r_k}\,dx\right.$$
$$+ \int_0^{\frac{\alpha-a_{k-1}}{1-a_{k-1}}} g_k\left(\frac{\alpha-x}{1-x}\right)x^{r_k-1}(1-x)^{s+1-r_k}\,dx$$
$$+ \sum_{d=k+1}^{\ell}\int_{I'_d} g_d(\alpha/x)x^{r_k}(1-x)^{s-r_k}\,dx$$
$$\left. + \sum_{d=1}^{k-1}\int_{I''_d} g_d\left(\frac{\alpha-x}{1-x}\right)x^{r_k-1}(1-x)^{s+1-r_k}\,dx\right],$$

is a contraction for all $1 \leq k \leq \ell$. Here g_k is the restriction of $g : [0,1] \to \mathbb{R}$ in the interval I_k, and I'_d and I''_d are defined as above. □

[2] We use $X \overset{\mathcal{L}}{=} Y$ to denote that the random variable X has the same distribution as the random variable Y; similarly, $X_n \xrightarrow{\mathcal{L}} X$ denotes convergence in law of the sequence of random variables $\{X_n\}_{n\geq 0}$ to the random variable X.

The technical details of the full proof of Theorem 1—which we have partially given here— are involved. But the intuition behind the statement of the theorem is rather simple: divide both sides of (2) by n, take $j = x \cdot n$ in the summations, and since we anticipate that $C_{n,m} \sim f(\alpha) \cdot n$, we 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

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

Passing to the limit when $n \to \infty$, the summations become integrals thus yielding the equations stated in the theorem. Since r is an integer function, the range of α must be broken into ℓ intervals, giving the piecewise definition of $f(\alpha)$.

Bearing in mind that discontinuities in the definition of $r(\alpha)$ carry on to discontinuities in $f(\alpha)$ we can also use the following simpler integral equation satisfied by the characteristic function $f(\alpha)$:

$$f(\alpha) = 1 + \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \left[\int_{\alpha}^{1} f(\alpha/x) x^{r(\alpha)} (1-x)^{s-r(\alpha)} \, dx \right.$$
$$\left. + \int_{0}^{\alpha} f\left(\frac{\alpha - x}{1-x}\right) x^{r(\alpha)-1}(1-x)^{s+1-r(\alpha)} \, dx \right]. \tag{3}$$

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 $k < \ell/2$; furthermore if the number of intervals ℓ is even then $a_{\ell/2} = 1/2$.

We immediately obtain the next lemma.

Lemma 1. *For any symmetric adaptive sampling strategy, its characteristic function is symmetric, i.e., $f(\alpha) = f(1-\alpha)$. More precisely, $f_k(\alpha) = f_{\ell+1-k}(1-\alpha)$, if $\alpha \in I_k$.*

Proof. Since $r(\alpha)$ is symmetric it is not difficult to prove that for any n and any m, $C_{n,m} = C_{n,n+1-m}$. The statement of the lemma immediately follows. □

Furthermore, we can prove the following lemma.

Lemma 2. *For any adaptive sampling strategy*

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

where $r_0 = \lim_{\alpha \to 0} r(\alpha)$ and all limits of $\alpha \to 0$ are taken from the right.

Proof. Taking the limit when $\alpha \to 0$ from the right in (3) and since $f(\alpha)$ is bounded in $[0,1]$, we have

$$\begin{aligned}
\lim_{\alpha\to 0} f(\alpha) &= 1 + \frac{s!}{(r_0-1)!(s-r_0)!}\int_0^1 \left(\lim_{\alpha\to 0} f(\alpha/x)\right) x^{r_0}(1-x)^{s-r_0}\,dx \\
&= 1 + \frac{s!}{(r_0-1)!(s-r_0)!}\int_0^1 \left(\lim_{\alpha\to 0} f(\alpha)\right) x^{r_0}(1-x)^{s-r_0}\,dx \\
&= 1 + \frac{s!}{(r_0-1)!(s-r_0)!}\cdot\left(\lim_{\alpha\to 0} f(\alpha)\right)\cdot\int_0^1 x^{r_0}(1-x)^{s-r_0}\,dx \\
&= 1 + \frac{s!}{(r_0-1)!(s-r_0)!}\cdot\left(\lim_{\alpha\to 0} f(\alpha)\right)\cdot\frac{r_0!(s-r_0)!}{(s+1)!}.
\end{aligned}$$

Hence,

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

□

From Lemma 2 we can easily rederive the known fact 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 proportion-from-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 for $s \geq 2$. In particular, proportion-from-s and similar strategies perform better than median-of-$(2t+1)$ variants, at least for low and high values of α. One of the main goals of this work is to establish how and when this happens.

To prepare for that journey we need a couple of additional technical results that are proved in Appendix A. In order to provide an explicit solution of the integral equations in Theorem 1, we transform, after lengthy and careful computations, the original problem to one of ordinary differential equations.

Lemma 3. *For any adaptive sampling strategy,*

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

where $f(\alpha)$ is its characteristic function, and $\alpha \in I_k$, $1 \leq k \leq \ell$.

Since proportion-from-s is a symmetric strategy, we only have to consider the equations for $1 \leq k \leq \lceil s/2 \rceil$ 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 $1 \leq k \leq \lceil s/2 \rceil$, since $r_k = k$,

$$\frac{d^{s+1-k}\phi_k}{dx^{s+1-k}} - \frac{s!}{(s-k)!}\frac{1}{(1-x)^k}\frac{d^{s+1-2k}\phi_k}{dx^{s+1-2k}} - \frac{s!}{(k-1)!}\frac{(-1)^{s+1-k}}{x^{s+1-k}}\phi_k(x) = 0. \qquad (4)$$

An important special case of the ordinary differential equation (ODE) above is for the central interval ($k = t+1$) when $s = 2t+1$. Then the ODE is identical to the corresponding ODE for median-of-$(2t+1)$.

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 $f_k(a_{k-1})$, $f_k'(a_{k-1}), \ldots, \frac{d^s f_k}{d\alpha^s}(a_{k-1})$ for every k, $1 \le k \le \ell$. In order to overcome this problem, we use a different technique, namely, substitute the f_k's in the integral equations by the general form of the solution for the corresponding differential equation and fix the values of the unknown constants by equalizing both sides. When the adaptive strategy is symmetric the problem is somewhat simpler because there are less f_k's to cope with and the argument of symmetry can be of help when determining the constants. However, the essential obstacles remain.

Last but not least the following result allows us to investigate the behavior of $C_n = \frac{1}{n}\sum_{1 \le m \le n} C_{n,m}$ as $n \to \infty$.

Lemma 4. *Let C_n be the average cost to select an element of random rank out of n elements using a symmetric adaptive sampling strategy. Then we have*

$$\overline{f} = \lim_{n\to\infty} \frac{C_n}{n} = \int_0^1 f(\alpha)\, d\alpha,$$

where $f(\alpha)$ is the characteristic function of the algorithm, and it is as given by Theorem 1.

3 Proportion-from-2

Let us begin with the simplest "proportion-from" strategy: $s = 2$. Solving (4) is not very difficult and even can be said to be routine in this case. 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 (4) is then

$$\frac{d^2\phi_1}{dx^2} - \frac{2}{1-x}\frac{d\phi_1}{dx} - \frac{2}{x^2}\phi_1(x) = 0, \tag{5}$$

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 $\lambda = -1$ and $\lambda = 2$. This entails a solution of the form $\phi_1(x) = \phi_{1,1}(x) + \phi_{1,2}(x)$ (see [20] and pages 14 to 15 in Section 4) where

$$\phi_{1,1}(x) = \sum_{n\ge 0} a_n x^{n-1} + A \cdot \phi_{1,2}(x) \cdot \ln x,$$
$$\phi_{1,2}(x) = \sum_{n\ge 0} b_n x^{n+2},$$

for some coefficients $\{a_n\}_{n\ge 0}$ and $\{b_n\}_{n\ge 0}$ and some constant A. The second term in $\phi_{1,1}(x)$ is necessary since the roots of the indicial equation differ by an integer constant.

Substituting the proposed form for $\phi_1(x)$ into the differential equation we obtain recurrence relations for the coefficients a_n and b_n, and from there a simple form for ϕ_1 which depends on two constants, since we can prove that $A = 0$, and $a_n = a_0$ and $b_n = b_0$ for all $n \geq 0$. Indeed,

$$\phi_1(x) = \frac{a_0}{x(1-x)} + \frac{b_0 x^2}{1-x}.$$

Integrating ϕ_1 twice we 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 constants. The painful process is to substitute the general expression for f_1 and f_2 into the integral equation (Theorem 1) and equalize the coefficients of powers of x in both sides.

Finally, one gets

$$a = -\frac{1}{12(3\ln 2 - 2)}, \qquad b = -2, \qquad c = -\frac{4\ln 2 - 3}{8(3\ln 2 - 2)}, \qquad d = -1/2.$$

The maximum of $f(\alpha)$ is at $\alpha = 1/2$. Indeed, since

$$f'(\alpha) = \frac{(48\ln 2 - 36)(\ln(1-\alpha) - \ln\alpha) - 4\ln\alpha - 2\alpha^2 - 4\alpha + 3}{8(3\ln 2 - 2)},$$

we have that $f'(0) = \infty$, $f'(1/2) > 0$ and since $f'(\alpha)$ is strictly decreasing, we conclude that it is 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[3].

We also have

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

which tells us that proportion-from-2 makes roughly $2.6n$ comparisons on average. Compare with the $3n$ comparisons of standard quickeslect and the $2.5n$ comparisons made by median-of-3.

It is interesting to notice that, as we expected, $f(\alpha) \leq m_0(\alpha)$ for all $0 \leq \alpha \leq 1$. Compared to median-of-3, proportion-from-2 is better when $\alpha \leq 0.140\ldots$ and, symmetrically, when $\alpha \geq 0.859\ldots$; and it is worse otherwise. The fact that we can outperform median-of-three with only two elements per sample is encouraging (see Figure 2). In that percentile, both algorithms make in average approximately $2.362\ldots n$ comparisons. However, it is a bit unfair to compare median-of-3 and proportion-from-2 since these strategies use a different number of elements in the samples (and standard quickselect uses samples of size $s = 1$).

[3] We only give four figures in the numerical values in this section and the following. Nevertheless 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.

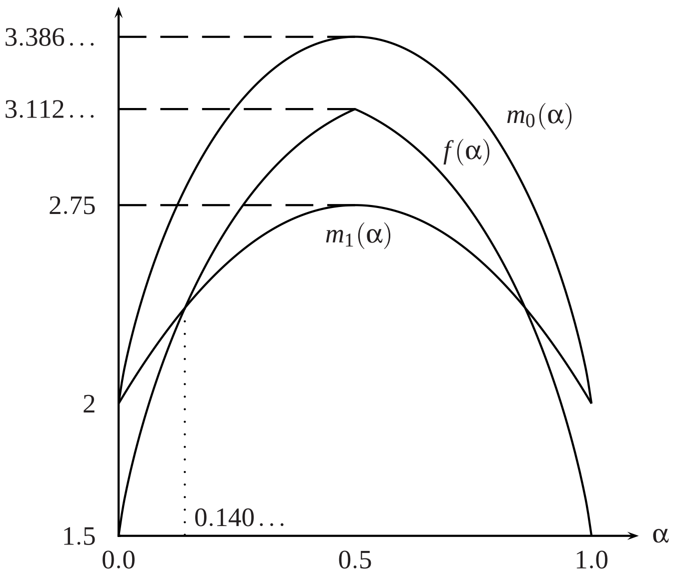


Figure 2: Plot of $m_0(\alpha)$, $m_1(\alpha)$ and the characteristic function $f(\alpha)$ of proportion-from-2.

4 Batfind: Proportion-from-3

The steps needed to analyze *batfind* (a.k.a. proportion-from-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,

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

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. The two indicial equations are $\lambda(\lambda-1)(\lambda-2)+6 = 0$, with roots -1, $2+i\sqrt{2}$ and $2-i\sqrt{2}$, and $\lambda(\lambda-1)-6 = 0$ with roots $\lambda = -2$ and $\lambda = 3$, respectively.

To solve the differential equation for $\phi_1(x)$ we use the identity $x^{i\sqrt{2}} = e^{i\sqrt{2}\ln x} =$

$\cos(\sqrt{2}\ln x)+i\sin(\sqrt{2}\ln x)$, and so we assume $\phi_1(x)=\phi_{1,1}(x)+\phi_{1,2}(x)+\phi_{1,3}(x)$ where

$$\begin{aligned}
\phi_{1,1}(x) &= \sum_{n\geq 0} a_n x^{n-1},\\
\phi_{1,2}(x) &= x^2\cos(\sqrt{2}\ln x)\sum_{n\geq 0} b_n x^n,\\
\phi_{1,3}(x) &= x^2\sin(\sqrt{2}\ln x)\sum_{n\geq 0} c_n x^n,
\end{aligned}$$

for some unknown coefficients $\{a_n\}_{n\geq 0}$, $\{b_n\}_{n\geq 0}$ and $\{c_n\}_{n\geq 0}$. Also, similarly to proportion-from-2, we propose $\phi_2(x)=\phi_{2,1}(x)+\phi_{2,2}(x)$, with

$$\phi_{2,1}(x)=\sum_{n\geq 0}\hat{a}_n x^{n-2}+A\cdot\phi_{2,2}(x)\cdot\ln x, \qquad \phi_{2,2}(x)=\sum_{n\geq 0}\hat{b}_n x^{n+3},$$

for some other coefficients $\{\hat{a}_n\}_{n\geq 0}$ and $\{\hat{b}_n\}_{n\geq 0}$ and some constant A.

Substituting the proposed form for $\phi_1(x)$ into the differential equation we obtain recurrence relations for the unknown coefficients $\{a_n\}_{n\geq 0}$, $\{b_n\}_{n\geq 0}$ and $\{c_n\}_{n\geq 0}$ and we finally get

$$\begin{aligned}
\phi_{1,1}(x) &= a_0\frac{1}{x(1-x)},\\
\phi_{1,2}(x) &= b_0\frac{x^2}{1-x}\cos(\sqrt{2}\ln x),\\
\phi_{1,3}(x) &= c_0\frac{x^2}{1-x}\sin(\sqrt{2}\ln x),
\end{aligned}$$

since $a_n=a_0$, $b_n=b_0$ and $c_n=c_0$ for all $n\geq 0$, for some arbitrary constants a_0, b_0 and c_0. Integrating ϕ_1 twice yields

$$f_1(x)=-C_0(1+\mathcal{H}(x))+C_1+C_2x+C_3\cdot K_1(x)+C_4\cdot K_2(x)$$

where

$$\begin{aligned}
K_1(x) &= \cos\left(\sqrt{2}\ln x\right)\cdot\sum_{n\geq 0}A_n x^{n+4}+\sin\left(\sqrt{2}\ln x\right)\cdot\sum_{n\geq 0}B_n x^{n+4},\\
K_2(x) &= \sin\left(\sqrt{2}\ln x\right)\cdot\sum_{n\geq 0}A_n x^{n+4}-\cos\left(\sqrt{2}\ln x\right)\cdot\sum_{n\geq 0}B_n x^{n+4},\\
A_n &= \frac{(n+2)(n+5)}{(n^2+6n+11)(n^2+8n+18)}, \quad B_n=\frac{\sqrt{2}(2n+7)}{(n^2+6n+11)(n^2+8n+18)}.
\end{aligned}$$

We can also determine in a similar way the values $\{\hat{a}_n\}$ and $\{\hat{b}_n\}$ up to arbitrary constants by substituting the proposed form for ϕ_2 into the corresponding differential equation yielding

$$\phi_{2,1}(x)=\hat{a}_0\left(\frac{1}{x^2}+\frac{1}{(1-x)^2}\right), \qquad \phi_{2,2}(x)=\hat{b}_0\frac{x^3(5x^3-20x^2+28x-14)}{14(1-x)^2}.$$

Then, integrating ϕ_2 three times and taking into account the symmetry of f_2, we get

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

The value of the constants C_i in $f_1(x)$ and $f_2(x)$ can be obtained by the same routine but cumbersome procedure of substitution into the integral equations that we have already used to analyze proportion-from-2. Thus we get that the constants C_i are the solutions to a system of equations which can be found in Appendix C. The coefficients given there are for the generalizations studied in Sections 5 and 6; the coefficients that we need here can be obtained by setting $\nu = 1/3$, $\xi_1 = 1$ and $\xi_2 = 0$ in the formulæ of Appendix C. We get then

$$\begin{aligned}&C_0 = -24/11, C_1 = -28/33, C_2 \doteq 0.193\ldots, C_3 \doteq -100.190\ldots,\\&C_4 \doteq -27.556\ldots, C_5 \doteq -1.463\ldots, C_6 \doteq 0.439\ldots, C_7 \doteq 0.135\ldots.\end{aligned}$$

The solution just obtained for batfind's characteristic function is quite representative of the general situation. For general s, the indicial equation of (4) is

$$\lambda^{\underline{s+1-k}} - (-1)^{s+1-k} s^{\underline{s+1-k}} = 0, \tag{7}$$

where $x^{\underline{k}} = x \cdot (x-1) \cdots (x-k+1)$ denotes the kth falling power of x, for any $k \geq 0$ [6].

If we denote its roots $\lambda_1, \ldots, \lambda_{s+1-k}$ in ascending order of their real part, we have $\lambda_1 = -k$ and then we have $\lfloor (s-k)/2 \rfloor$ pairs of complex conjugate roots. If $s+1-k$ is odd then there are no more roots, but if $s+1-k$ is even then $\lambda_{s+1-k} = s$ is also a root. All the roots have their real parts between $-k$ and s and, except for the case $\lambda_1 = -k$ and $\lambda_{s+1-k} = s$ when $s+1-k$ is even, no pair of roots has an integer difference. The proofs of these facts are more or less involved and are similar in spirit to those of Mahmoud and Pittel [16] in their analysis of the space of m-ary search trees.

Then, for $\phi_k^{(s)} = d^{k+1} f_k^{(s)} / dx^{k+1}$, which satisfies the linear differential equation (4), we have a solution of the form

$$\phi_k^{(s)}(x) = \phi_{k,1}^{(s)}(x) + \cdots + \phi_{k,s+1-k}^{(s)}(x),$$

with

$$\phi_{k,j}^{(s)}(x) = \sum_{n \geq 0} a_n^{(j)} x^{n+\lambda_j}, \tag{8}$$

for some coefficients $\{a_n^{(j)}\}_{n \geq 0}$. However, if $s+1-k$ is even then $\lambda_1 = -k$ and $\lambda_{s+1-k} = s$, the difference of this pair of roots is an integer, and hence a slightly different form for $\phi_{k,1}^{(s)}(x)$ must be assumed, namely,

$$\phi_{k,1}^{(s)}(x) = \sum_{n \geq 0} a_n^{(1)} x^{n-k} + A \cdot \phi_{k,s+1-k}^{(s)}(x) \cdot \ln x. \tag{9}$$

It is important to point out that the roots λ_j depend on s and k and the coefficients $\{a_n^{(j)}\}$ depend on s and k as well; but we have refrained to use additional sub- or superscripts to make explicit that dependence.

Substituting the general form $\phi_k^{(s)}(x)$ back into the differential equation (4) and equalizing on powers of x yields the recurrences satisfied by the coefficients $a_n^{(j)}$. However, we were able to find explicit solutions of these recurrences only for some special cases. Another outcome of our partial analysis is that the logarithmic extra term in $\phi_{k,1}^{(s)}(x)$ when $s+1-k$ is even (see Equation 9) actually vanishes, since it can be proved that $A=0$.

Altogether, by integrating $k+1$ times, this leads to the general solution of the form

$$\begin{aligned} f_k^{(s)}(x) &= -C_{k+1}(1+\mathcal{H}(x)) + C_k x^k + C_{k-1}x^{k-1} + \cdots + C_0 \\ &\quad + \sum_{j=2}^{s+1-k} C_{k+j} x^{\lambda_j+k+1} \sum_{n\geq 0} \frac{a_n^{(j)}}{(n+k+1+\lambda_j)^{\underline{k+1}}} x^n, \end{aligned}$$

for some arbitrary constants $C_0, \ldots, C_{s+1}$ and coefficients $\{a_n^{(j)}\}$. For each k we have a different set of $s+2$ arbitrary constants, and the coeffients $\{a_n^{(j)}\}$ depend on both k and s. Furthermore, as we have $\lfloor (s-k)/2 \rfloor$ pairs of complex conjugate roots $\lambda_j = \mu_j \pm \tau_j \mathbf{i}$, we may write for $s+1-k$ odd

$$\begin{aligned} f_k^{(s)}(x) &= -C_{k+1}(1+\mathcal{H}(x)) + C_k x^k + C_{k-1}x^{k-1} + \cdots + C_0 \\ &\quad + \sum_{j=1}^{(s-k)/2} C_{2j+k} x^{\mu_{2j}+k+1} K_1^{(j)}(x) + C_{2j+k+1} x^{\mu_{2j}+k+1} K_2^{(j)}(x), \end{aligned} \tag{10}$$

with

$$\begin{aligned} K_1^{(j)}(x) &= \cos(\tau_{2j}\log x) \sum_{n\geq 0} A_n^{(j)} x^n + \sin(\tau_{2j}\log x) \sum_{n\geq 0} B_n^{(j)} x^n \\ K_2^{(j)}(x) &= \sin(\tau_{2j}\log x) \sum_{n\geq 0} A_n^{(j)} x^n - \cos(\tau_{2j}\log x) \sum_{n\geq 0} B_n^{(j)} x^n, \end{aligned}$$

and

$$\begin{aligned} A_n^{(j)} &= \frac{\Re\left(a_n^{(2j)} (n+k+1+\lambda_{2j})^{\underline{k+1}}\right)}{\prod_{i=0}^k \left((n+k+1+\mu_{2j}-i)^2 + \tau_{2j}^2 \right)}, \\ B_n^{(j)} &= \frac{\Im\left(a_n^{(2j)} (n+k+1+\lambda_{2j})^{\underline{k+1}}\right)}{\prod_{i=0}^k \left((n+k+1+\mu_{2j}-i)^2 + \tau_{2j}^2 \right)}. \end{aligned}$$

The same holds for $s+1-k$ even, but we have to add to (10) the additional term

$$C_{s+1} x^{s+k+1} \sum_{n\geq 0} \frac{a_n^{(s+1-k)}}{(n+k+1+s)^{\underline{k+1}}} x^n,$$

corresponding to the root $\lambda_{s+1-k} = s$.

Coming back to batfind, 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

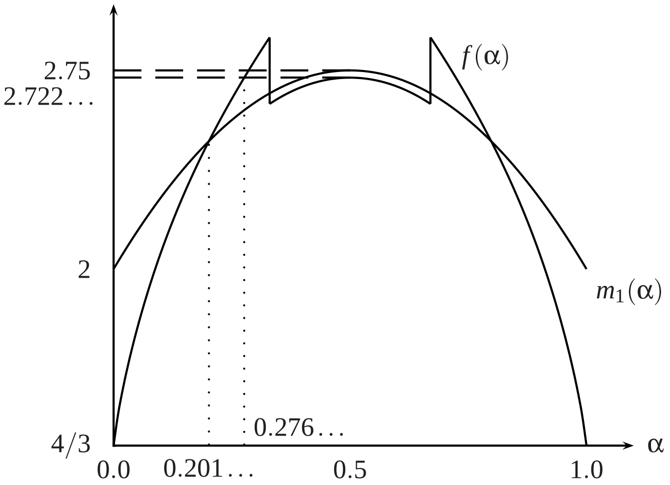


Figure 3: Plot of batfind's characteristic function $f(\alpha)$ and $m_1(\alpha)$.

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 $1/3 < \alpha < 2/3$, and the same is true for the intervals $\alpha \leq 0.201\ldots$ and $\alpha \geq 0.798\ldots$; see Figure 3. In particular, $f(1/2) = f_2(1/2) \doteq 2.722\ldots < m_1(1/2) = 2.75$. However, it came as a surprise the fact that $\alpha = 1/2$ is not the most difficult relative rank for batfind: for instance, $f(1/3) = f_1(1/3) \doteq 2.883\ldots > f(1/2)$.

Finally, the integration of $f(\alpha)$ 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 by 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 for these facts is the following. Assume for the sake of concreteness that $n = 1000$ and $m = 332$. While there is some chance (in particular $\sim 29.6\%$ of the times) that the rank of the pivot selected by batfind is close but larger than $1000/3 = 333.3$ and then we discard almost two thirds of the input, it is more likely for the rank of the pivot to be less than m and then we discard a bit less than a third of the input (this happens around 70.4% of the times). In the latter case, at the next recursive call, the rank of the sought element would 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 or equal to $n/3$. In other words, this means that to find an element of rank α

smaller than but close to 1/3 we should have chosen the median and not the smallest element as pivot.

5 ν-find: A variant of batfind

The natural question raised in the previous section is: for which values of α should we pick the smallest, the median or the largest element of the sample? From the lessons of Section 4, it seems clear that the median of the sample must be used for a larger interval of α. However, if we make the central interval too large we may lose all the benefits of proportion-from-3 sampling. Altogether, this suggests that there should exist an optimal choice for the endpoints of the intervals. The main goal of this section is to prove this assertion (Theorem 2).

One important point is that no matter how we choose the endpoints, the corresponding characteristic function $f(\alpha)$ satisfies the differential equations (6) of Section 4. Hence, the general form of the f_k's is exactly the same as before and the only difference is in the value of the involved constants, because of the different initial conditions.

More formally, the goal of this section is to investigate the properties of the characteristic function 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 characteristic function corresponding to this strategy, which we call ν-find. When $\nu \to 0$, ν-find behaves as quickselect with median-of-three. When $\nu = 1/3$, we have batfind. Finally, when $\nu \to 1/2$, the median of the sample is always discarded, so ν-find behaves slightly different than proportion-from-2. This "pseudo-proportion-from-2" variant is not interesting at all; it does even worse than proportion-from-2 in a large interval of α.

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 investigate the dependence on ν of the values of the constants C_i's in the general form of $f_{1,\nu}$ and $f_{2,\nu}$ (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, $C_6(\nu) = 7/4 \cdot C_5(\nu) + 3$. Appendix C provides the values of the remaining C_i's as functions of ν. Actually, we give there the values for the C_i's corresponding to the average total cost (see Section 6); the values of the C_i's corresponding to the average number of comparisons can easily be obtained by setting $\xi_1 = 1, \xi_2 = 0$ in the given formulæ.

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). It is also important to point out that for fixed ν, $\lim_{\alpha\to 0} f_\nu(\alpha) = \lim_{\alpha\to 0} f_{1,\nu}(\alpha) = 4/3$. However, $\lim_{\nu\to 0} f_{1,\nu}(\nu) = 3/2$ and $\lim_{\nu\to 0} f_{2,\nu}(\nu) = 2$.

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

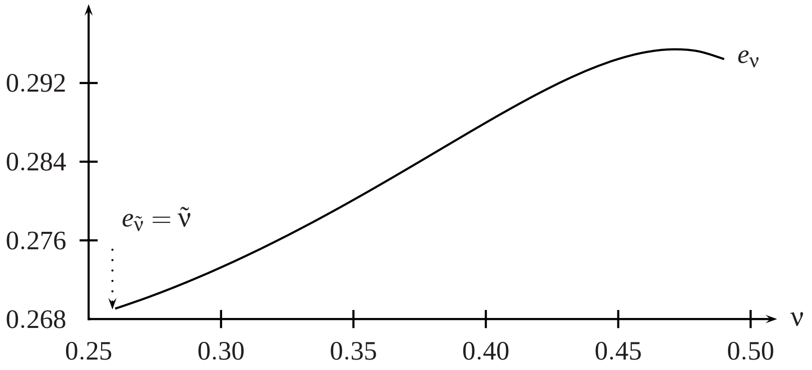


Figure 4: Plot of e_ν.

$\alpha = 1-\nu$. So $\tilde{\nu}$ is the solution of $f_{1,\nu}(\nu) = f_{2,\nu}(1/2)$. Numerical computations yield $\tilde{\nu} \doteq 0.268\ldots$. This phenomenon leads naturally to the concept of *expensive* ranks. We say that $\alpha \neq 1/2$ is expensive if $f_\nu(\alpha) \geq f_\nu(1/2)$. If $\nu < \tilde{\nu}$ then there are no expensive ranks. But if $\nu \geq \tilde{\nu} \doteq 0.268\ldots$ then $[e_\nu, \nu]$ and $[1-\nu, 1-e_\nu]$ are the intervals containing the expensive ranks. To compute e_ν we just need to solve $f_{1,\nu}(e_\nu) - f_\nu(1/2) = 0$ (see Figure 4).

5.1 The optimal value of ν

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 Figures 5 and 6). 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 that we state in the main theorem of this section.

Theorem 2. *There exists an optimal value of* ν, *namely* $\nu^* \doteq 0.182\ldots$, *such that* $f_{1,\nu^*}(\nu^*) = f_{2,\nu^*}(\nu^*)$ *and for all* ν, $0 < \nu < 1/2$, *and for all* α, $0 \leq \alpha \leq 1$,

$$f_{\nu^*}(\alpha) \leq f_\nu(\alpha).$$

Despite the technical difficulties of the proof (given in Appendix B), the intuitive explanation for Theorem 2 is easy: if $\nu < \nu^*$ then $f_{1,\nu}(\nu) < f_{2,\nu}(\nu)$, which means that for some values of $\alpha > \nu$ close enough to ν we would be doing better by choosing the smallest element in the sample rather than the median; on the contrary, if $\nu > \nu^*$ then $f_{1,\nu}(\nu) > f_{2,\nu}(\nu)$, and that means that for some $\alpha \leq \nu$ the algorithm should have chosen (as in batfind) the median as the pivot, not the smallest. At $\nu = \nu^*$ we are just choosing the right pivot for each relative rank α.

Since ν^* optimizes f_ν it minimizes the maxima; in particular, ν^* minimizes $f_\nu(1/2)$. Also, since $\nu^* < \tilde{\nu}$, it follows that $\alpha = 1/2$ is the most difficult relative rank for ν^*-find, where we have $f_{\nu^*}(1/2) \doteq 2.659\ldots$ (see Figure 7).

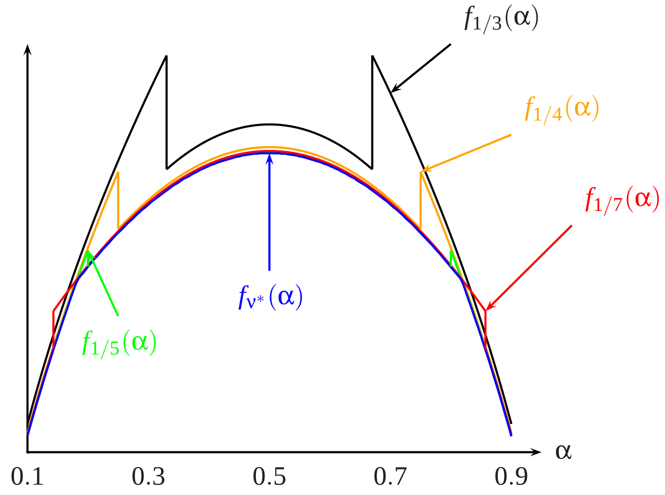


Figure 5: Plot of $f_\nu(\alpha)$ for $\nu \in \{1/3, 1/4, 1/5, \nu^*, 1/7\}$.

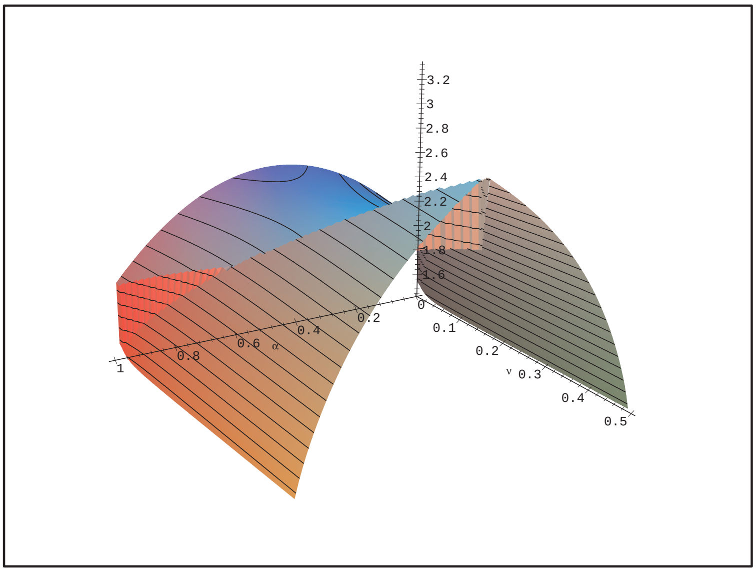


Figure 6: 3-dimensional plot for ν-find.

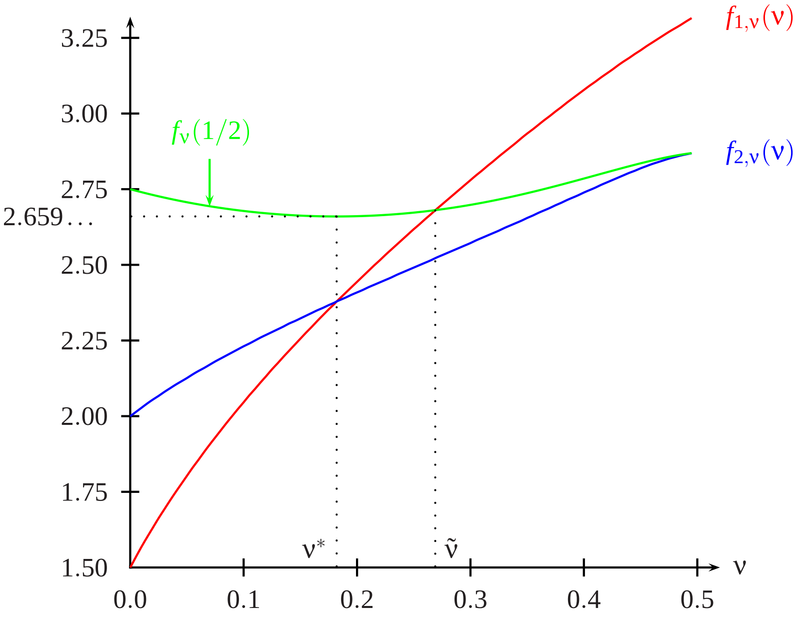


Figure 7: Plot of $f_{1,\nu}(\nu)$, $f_{2,\nu}(\nu)$ and $f_\nu(1/2)$.

It is also obvious that ν^* must minimize the average value $\overline{f}_\nu$. In particular, we have $\overline{f}_{\nu^*} \doteq 2.342\ldots$ (see Figure 8). And ν^*-find must outperform median-of-three as $m_1(\alpha) = \lim_{\nu\to 0} f_\nu(\alpha)$.

It is reasonable to think that when using samples of size s, a suitable choice $a_1^*, a_2^*, \ldots$ of the interval endpoints that makes $f^{(s)}(\alpha)$ continuous is the optimal choice for each possible s, like we have just proved for $s = 3$.

Conjecture 1. *Fix some value $s \geq 3$ and let*

$$\mathcal{A}_s = \{(0, a_1, a_2, \ldots, a_{s-1}, 1) \,|\, 0 < a_1 < a_2 < \cdots < a_{s-1} < 1 \text{ and } \\ 1 - a_{s-k} = a_k \text{ for all } k, 1 \leq k \leq \lfloor s/2 \rfloor\}.$$

Let $f_{\mathbf{a}}(\alpha)$ denote the function corresponding to proportion-from-s sampling with endpoints at $\mathbf{a} = (0, a_1, a_2, \ldots, a_{s-1}, 1) \in \mathcal{A}_s$. Then, there exists a unique $\mathbf{a}^* \in \mathcal{A}_s$ such that $f^{(s)}_{\mathbf{a}^*}(\alpha)$ is continuous for $\alpha \in [0,1]$ and

$$f^{(s)}_{\mathbf{a}^*}(\alpha) \leq f^{(s)}_{\mathbf{a}}(\alpha),$$

for any $\mathbf{a} \in \mathcal{A}_s$ and any $0 \leq \alpha \leq 1$.

This conjecture is the analogous of Theorem 2 for general s. Even though we have been unable to complete all the technical details needed to prove this conjecture, all the evidence indicates that such an optimal choice of the endpoints of the intervals must exist. As noticed in Section 4 for proportion-from-3, proportion-from-s algorithms

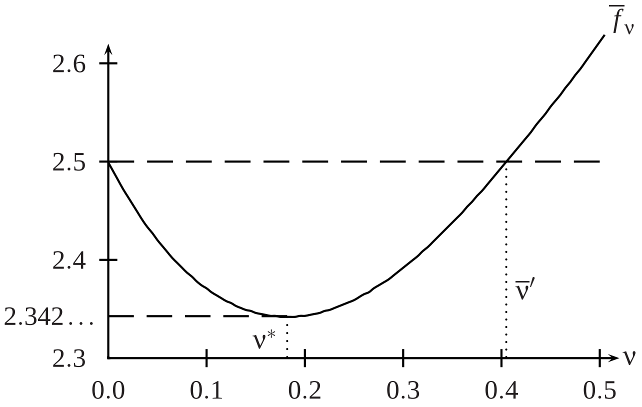


Figure 8: Plot of $\overline{f}_\nu$; $\overline{m}_1 = 2.5$ is also depicted for convenience.

also have discontinuities near the endpoints of the intervals. This problem is originated by a "bad choice" of the pivot for ranks close to the interval endpoints. Following the arguments of this section, we can expect the existence of an optimal choice of the values $a_1 \cdots a_{s-1}$ such that the function $f^{(s)}_{\mathbf{a}}(\alpha)$ is continuous and minimum for all α. Also, since median-of-$(2t+1)$ is the special case of proportion-from$(2t+1)$ when $a_1 \to 0, a_2 \to 0, \ldots, a_t \to 0, a_{t+1} \to 1, \ldots, a_{2t} \to 1$, then, if Conjecture 1 holds, it follows that $f^{(2t+1)}_{\mathbf{a}^*}$ must outperform median-of-$(2t+1)$ for any t.

5.2 A comparative study of ν-find

In order to compare ν-find with other algorithms, we use the fact that, except at $\nu = \alpha$, the function $f_\nu(\alpha)$ is continuous in ν for fixed α. Furthermore, the function has a local minimum at $\nu = \nu^*$ and its second derivative (except at $\nu = \alpha$) is strictly positive. Therefore, if we want to compare an algorithm G whose characteristic function is $g(x)$ with ν-find, it is enough to compare $g(x)$ to $\max\{f_{1,x}(x), f_{2,x}(x)\}$. Whenever $g(x)$ is above $\max\{f_{1,x}(x), f_{2,x}(x)\}$, the corresponding ν-find beats G for all ranks α; if $g(x)$ is below $\max\{f_{1,x}(x), f_{2,x}(x)\}$ that means that G beats ν-find in some ranks.

When comparing ν-find with standard quickselect, the result is clear cut: ν-find beats this algorithm for any rank α and any value of ν (see Figure 9).

Things get more intriguing when we compare ν-find with proportion-from-2 and with median-of-three. There are ranges of ν where ν-find is not uniformly better than proportion-from-2. In particular if $\nu \leq \nu^{(2)}_1 \doteq 0.116\ldots$ or $\nu \geq \nu^{(2)}_2 \doteq 0.347\ldots$ then $f_\nu(\alpha) \geq f^{(2)}(\alpha)$ in some ranges of α (see Figure 9). For instance, we already knew that proportional-of-2 does better than median-of-3 (which is the limit of ν-find when $\nu \to 0$) when α is sufficiently close to 0 or to 1. The values $\nu^{(2)}_1$ and $\nu^{(2)}_2$ are the solutions to $f_{2,\nu}(\nu) - f^{(2)}(\nu) = 0$ and $f_{1,\nu}(\nu) - f^{(2)}(\nu) = 0$, respectively.

A set of interesting values of ν also arises when we compare ν-find and median-of-three. In particular:

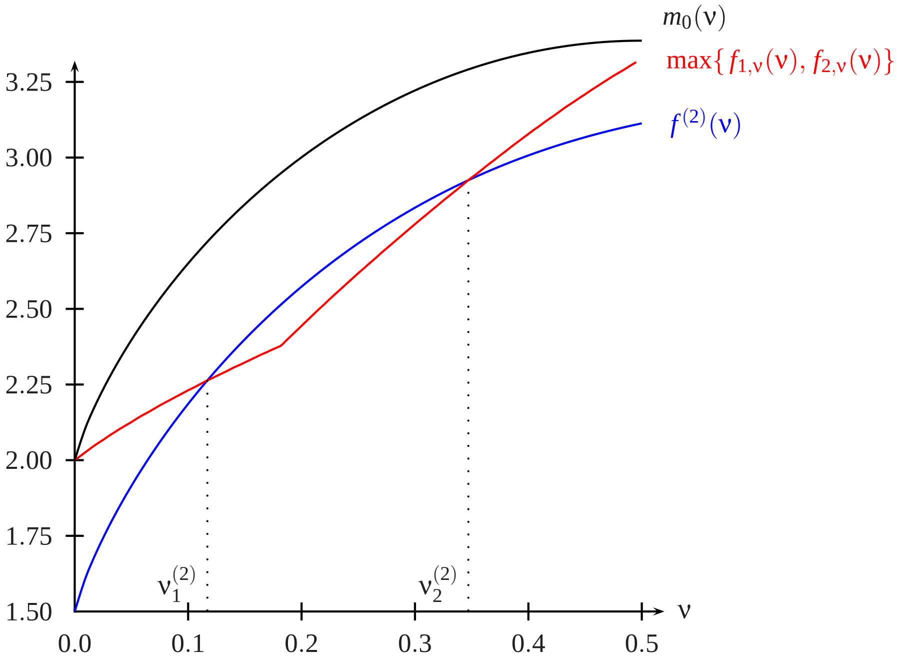


Figure 9: Plot of ν-find ($\max\{f_{1,\nu}(\nu), f_{2,\nu}(\nu)\}$) compared to standard quickselect ($m_0(\nu)$) and proportion-from-2 ($f^{(2)}(\nu)$).

1. For $\nu \leq \overline{\nu}' \doteq 0.404\ldots$, ν-find does better than median-of-3 on the average; that is, $\overline{f}_\nu \leq \overline{m}_1 = 5/2 = 2.5$ (see Figure 8).

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 10).

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 $\nu^* \leq \nu' \leq \tilde{\nu}$, hence when ν-find beats median-of-3, $\alpha = 1/2$ is already the most difficult relative rank for ν-find; but on the other hand, $f_{1,\nu'}(\nu') > f_{2,\nu'}(\nu')$ (see Figure 10).

4. If $\nu > \nu'$ then, by definition, ν-find does worse than median-of-three for some intervals of α. In particular, if $\nu' < \nu \leq \nu'_m$ then ν-find beats median-of-3 in $[0, \alpha_\nu]$, $(\nu, 1-\nu)$ and $[1-\alpha_\nu, 1]$; if $\nu'_m < \nu \leq \nu'' \doteq 0.381\ldots$ then ν-find beats median-of-3 in $[0, \alpha_\nu]$, $[1-\alpha_\nu, 1]$ and two subintervals of $(\nu, 1-\nu)$ not including $\alpha = 1/2$; finally, if $\nu > \nu''$ then ν-find beats median-of-three only in the intervals $[0, \alpha_\nu]$ and $[1-\alpha_\nu, 1]$. The value ν'' is the solution of the equation $f_{2,\nu}(\nu) - m_1(\nu) = 0$. 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 $m_1(\alpha_\nu) - f_{1,\nu}(\alpha_\nu) = 0$ (see Figure 11).

The value of f_ν at relevant points and related quantities, for several values of ν, are listed in Table 1.

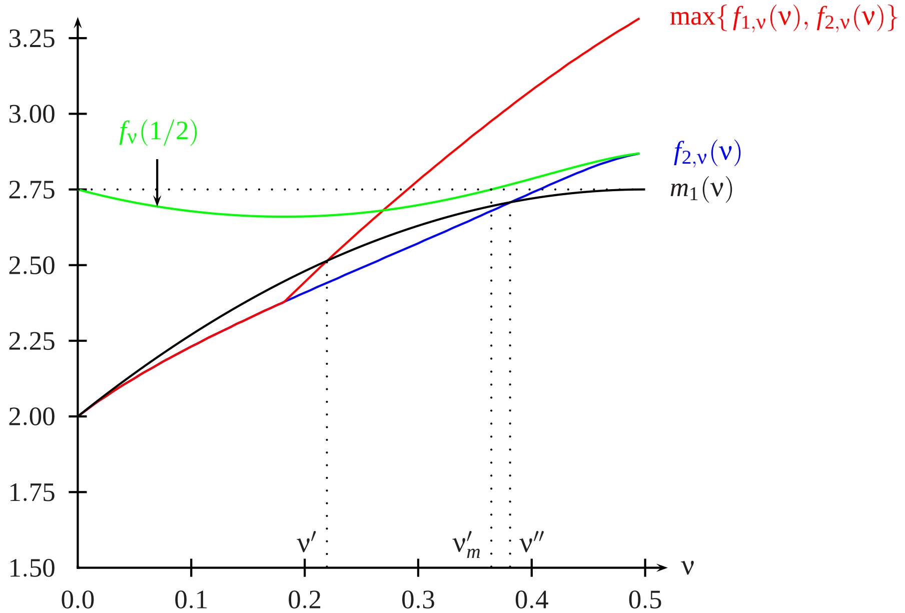


Figure 10: Plot of ν-find compared to median-of-three.

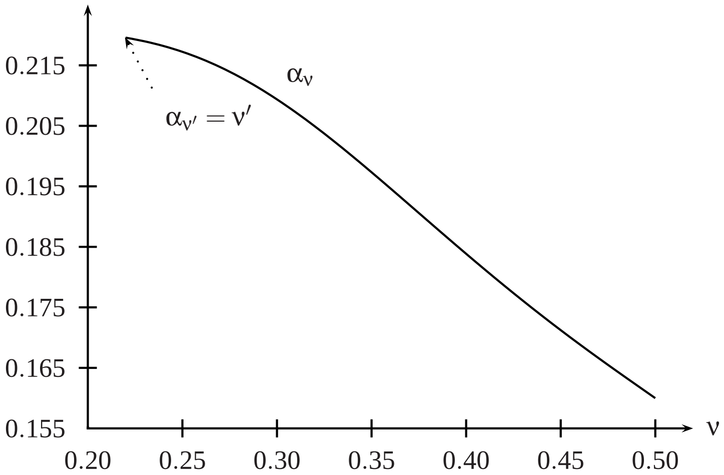


Figure 11: Plot of α_ν.

ν	$f_\nu(1/2)$	$f_{1,\nu}(\nu)$	$f_{2,\nu}(\nu)$	α_ν	e_ν	$\overline{f}_\nu$
$\nu \to 1/2$	$2.871\ldots$	$3.326\ldots$	$2.871\ldots$	$0.160\ldots$	$0.294\ldots$	$2.622\ldots$
$\overline{\nu}' \doteq 0.404\ldots$	$2.790\ldots$	$3.091\ldots$	$2.747\ldots$	$0.182\ldots$	$0.287\ldots$	2.5
$\nu'_m \doteq 0.364\ldots$	2.75	$2.976\ldots$	$2.679\ldots$	$0.193\ldots$	$0.280\ldots$	$2.453\ldots$
$1/3$	$2.722\ldots$	$2.883\ldots$	$2.627\ldots$	$0.201\ldots$	$0.276\ldots$	$2.421\ldots$
$\tilde{\nu} \doteq 0.268\ldots$	$2.680\ldots$	$2.680\ldots$	$2.522\ldots$	$0.214\ldots$	$\tilde{\nu}$	$2.370\ldots$
$1/4$	$2.672\ldots$	$2.617\ldots$	$2.491\ldots$	$0.217\ldots$	-	$2.359\ldots$
$\nu' \doteq 0.219\ldots$	$2.663\ldots$	$2.514\ldots$	$2.441\ldots$	ν'	-	$2.348\ldots$
$1/5$	$2.660\ldots$	$2.444\ldots$	$2.409\ldots$	-	-	$2.343\ldots$
$\hat{\nu} = \nu^*$ $\doteq 0.182\ldots$	$2.659\ldots$	$2.379\ldots$	$2.379\ldots$	-	-	$2.342\ldots$
$1/6$	$2.660\ldots$	$2.321\ldots$	$2.352\ldots$	-	-	$2.343\ldots$
$1/7$	$2.664\ldots$	$2.228\ldots$	$2.311\ldots$	-	-	$2.348\ldots$
$1/8$	$2.668\ldots$	$2.154\ldots$	$2.278\ldots$	-	-	$2.356\ldots$
$1/9$	$2.673\ldots$	$2.095\ldots$	$2.252\ldots$	-	-	$2.363\ldots$
$1/10$	$2.678\ldots$	$2.046\ldots$	$2.231\ldots$	-	-	$2.371\ldots$
$\nu \to 0$	2.75	$3/2$	2	-	-	2.5

Table 1: Some relevant parameters of ν-find.

In this section we have generalized the proportion-from-3 algorithm by allowing $a_1 = \nu$ to be any value in $(0, 1/2)$, instead of the arbitrary value $1/3$. In doing so, we have done a fine-tuning of the algorithm, finding the best choice for a_1 and taking care of the problems presented by batfind. Moreover, we were able to completely characterize the evolution of the algorithm as ν varies, and to prove the existence of the optimal value ν^* of a_1. Moreover ν^*-find outperforms median-of-3 in every range. Given the simplicity of its implementation, ν^*-find is a strong candidate for being the selection algorithm of choice in general-purpose libraries. Nevertheless, if we want to consider the practical impact of ν-find we should also study the number of exchanges made. This issue is studied in the next section.

6 Exchanges and total cost

An important part of the cost of the selection algorithm comes from the exchanges performed during the partition stages. It is thus interesting to consider the *average total cost* of the algorithm, where we define the total cost as a weighted sum of exchanges and comparisons. Other costs, such as the cost of the comparisons needed to select pivots or the cost of the bookkeeping associated to each iteration, can be neglected for our analysis since they are $o(n)$.

Taking into account exchanges introduces yet another twist in our framework. It is relatively easy to set up the integral equations to analyze the average total cost. However, a new difficulty arises here because the toll function depends now both on n and m. In particular, the average number of exchanges in a single partitioning step of an array of size n when we select according to the adaptive strategy given by $r(\alpha)$

is [18, 21]

$$\sum_{1\le j\le n} \pi^{(s,r)}_{n,j} \sum_t t \frac{\binom{j-1}{t}\binom{n-j}{t}}{\binom{n-1}{j}} = \frac{r(\alpha)(s+1-r(\alpha))}{(s+1)(s+2)} n + o(n).$$

Hence, if ξ_1 denotes the unit cost of a comparison, ξ_2 denotes the unit cost of an exchange and $X_{n,m}$ denotes the average number of exchanges made to select the mth element out of n elements, and we let $t(\alpha) = \lim_{n\to\infty, m/n\to\alpha}(\xi_1 \cdot C_{n,m} + \xi_2 \cdot X_{n,m})/n$, then we have $t(\alpha) = t_k(\alpha)$ if $\alpha \in I_k$, $1 \le k \le \ell$, and

$$\begin{aligned} t_k(\alpha) = \left(\xi_1 + \xi_2 \frac{r_k(s+1-r_k)}{(s+1)(s+2)}\right) + \frac{s!}{(r_k-1)!(s-r_k)!} \Bigg[& \\ \int_{\alpha/a_k}^{1} t_k(\alpha/x)\, x^{r_k}(1-x)^{s-r_k}\, dx & \\ + \int_0^{\frac{\alpha-a_{k-1}}{1-a_{k-1}}} t_k\left(\frac{\alpha-x}{1-x}\right) x^{r_k-1}(1-x)^{s+1-r_k}\, dx & \\ + \sum_{d=k+1}^{\ell} \int_{I'_d} t_d(\alpha/x)\, x^{r_k}(1-x)^{s-r_k}\, dx & \\ + \sum_{d=1}^{k-1} \int_{I''_d} t_d\left(\frac{\alpha-x}{1-x}\right) x^{r_k-1}(1-x)^{s+1-r_k}\, dx \Bigg], & \end{aligned}$$

with $I'_d = (\alpha/a_d, \alpha/a_{d-1})$ and $I''_d = \left(\frac{\alpha-a_d}{1-a_d}, \frac{\alpha-a_{d-1}}{1-a_{d-1}}\right)$. In particular, for ν-find, following the same steps of Section 2 we arrive at the same differential equations to be satisfied by the t_k's. Hence the general form of the t_k's is the same as for the f_k's but the involved constants are different. Once the corresponding C_i's have been determined (as functions of ν, ξ_1 and ξ_2) we can investigate the behavior of the average total cost and compare it to the other alternatives. Moreover, an analogue of Lemma 2 holds for the total cost of any adaptive sampling strategy: if $\lim_{\alpha\to 0} r(\alpha) = 1$ then $\lim_{\alpha\to 0} t(\alpha) = \xi_1 \cdot (1 + 1/s) + \xi_2/(s+2)$. In particular, for ν-find the average number of exchanges when $\alpha \to 0$ is $\lim_{\alpha\to 0} t_\nu(\alpha) = 1/5$ for any ν.

We can establish the existence of an optimal choice ν^* for the average total cost of ν-find, now depending on ξ_1 and ξ_2, which satisfies $t_{\nu^*}(\alpha) \le t_\nu(\alpha)$ for any value of ν and any α. Furthermore, we have also that ν^* makes t_ν continuous at $\alpha = \nu^*$. Figure 12 and Table 2 show the variation of ν^* as a function of $\xi = \xi_2/\xi_1$, which is actually the relevant parameter.

A particular interesting value of ν^* is for $\xi_1 = 0, \xi_2 = 1$, when we want to optimize the average number of exchanges. Then $\nu^* = \nu^*(\infty) \doteq 0.429\ldots$. For that optimum, the factor that multiplies n in the average number of exchanges to find the median is $t_{\nu^*}(1/2) \doteq 0.479\ldots$ and for the average number of exchanges to find an element of random rank we have $\bar{t}_{\nu^*} \doteq 0.391\ldots$.

Like in the previous section, we can define expensive ranks, and also determine those values of ν where the total cost of ν-find outperforms that of median-of-three on random ranks, to locate the median and to locate any rank. Similarly to the case of

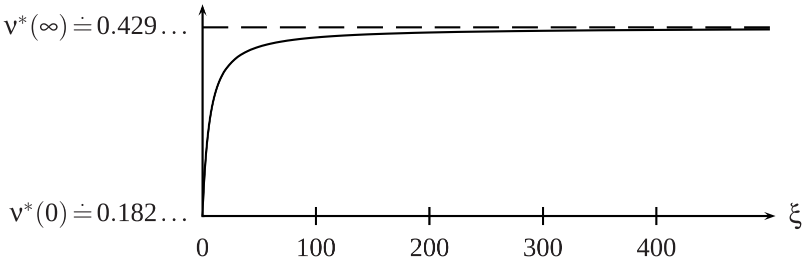


Figure 12: Plot of $\nu^* = \nu^*(\xi)$.

ξ	ν'	ν^*	ξ	ν'	ν^*
0	$0.219\ldots$	$0.182\ldots$	30	1/2	$0.389\ldots$
1	$0.270\ldots$	$0.213\ldots$	40	1/2	$0.398\ldots$
2	$0.313\ldots$	$0.239\ldots$	50	1/2	$0.403\ldots$
3	$0.349\ldots$	$0.259\ldots$	60	1/2	$0.407\ldots$
4	$0.380\ldots$	$0.276\ldots$	70	1/2	$0.410\ldots$
5	$0.405\ldots$	$0.290\ldots$	80	1/2	$0.413\ldots$
6	$0.427\ldots$	$0.302\ldots$	90	1/2	$0.414\ldots$
7	$0.445\ldots$	$0.312\ldots$	100	1/2	$0.416\ldots$
8	$0.461\ldots$	$0.321\ldots$	110	1/2	$0.417\ldots$
9	$0.474\ldots$	$0.328\ldots$	120	1/2	$0.418\ldots$
10	$0.486\ldots$	$0.335\ldots$	130	1/2	$0.419\ldots$
11	$0.497\ldots$	$0.341\ldots$	140	1/2	$0.419\ldots$
12	1/2	$0.346\ldots$	150	1/2	$0.420\ldots$
13	1/2	$0.351\ldots$	200	1/2	$0.422\ldots$
14	1/2	$0.355\ldots$	300	1/2	$0.424\ldots$
15	1/2	$0.358\ldots$	400	1/2	$0.426\ldots$
16	1/2	$0.362\ldots$	500	1/2	$0.426\ldots$
20	1/2	$0.373\ldots$	$\to \infty$	1/2	$0.429\ldots$

Table 2: Values of ν^* and ν' as functions of $\xi = \xi_2/\xi_1$.

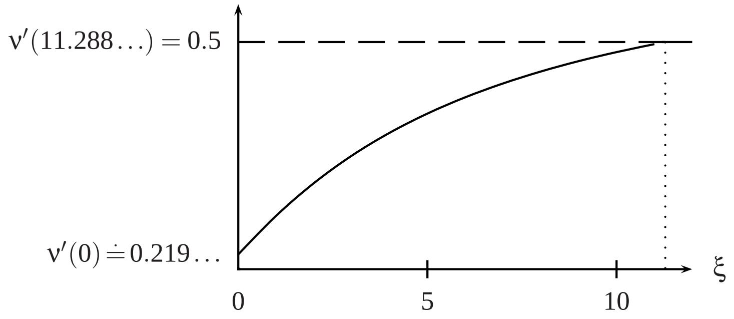


Figure 13: Plot of $\nu' = \nu'(\xi)$.

the optimal ν^*, all these special values of ν or α depend on $\xi = \xi_2/\xi_1$. For example, Figure 13 depicts the value ν' where the average total cost of ν-find is better than the average total cost of median-of-three on any relative rank α, as a function of $\xi = \xi_2/\xi_1$. The values of ν' are also given in Table 2. Of course, since ν^*-find is optimal, it must outperform median-of-3 on all ranks, so we have $\nu'(\xi) \geq \nu^*(\xi)$ for all ξ.

If $\xi \geq 11.288\ldots$ then $\nu' = 1/2$. In general ν' is given by the solution of $t_{2,0}(\nu) - t_{1,\nu}(\nu) = 0$, where $t_{2,0}(\alpha)$ is the characteristic function for the average total cost of median-of-three. When $\xi \geq 11.288\ldots$ the equation has no solution in $(0, 1/2)$; in other words, for any value of ν, ν-find beats median-of-3 on all ranks, so we assume by convention $\nu' = 1/2$. However, large values of ξ should not occur in practice; if the exchange of two elements were too expensive then we would handle an array of pointers to the elements instead.

It is not difficult to show that $t_{2,0}(\alpha) = (\xi_1 + \xi_2/5) \cdot m_1(\alpha) = (\xi_1 + \xi_2/5) \cdot (2 + 3\alpha(1-\alpha))$, i.e., the characteristic function $m_1(\alpha)$ for the average number of comparisons times the constant $(\xi_1 + \xi_2/5)$ that multiplies n in the toll function for the total cost recurrence. In general, the same is true for any adaptive sampling strategy which only defines one interval (standard quickselect, median-of-$(2t+1)$) or for proportion-from-2 because of the symmetry; the function $t(\alpha)$ for the average total cost is given by the characteristic function $f(\alpha)$ corresponding to comparisons times the factor that multiplies n in the toll function. This easily follows from the fact that if $C_{n,m}$ is the solution to recurrence (2) with toll function $n + o(n)$ then $\beta \cdot C_{n,m}$ is the solution to the recurrence with toll function $\beta \cdot n + o(n)$.

Last but not least, from a practical standpoint, if we take $\xi_1 = 4$ and $\xi_2 = 11$ as representative values for the cost of comparisons and exchanges (as suggested in [14]) then $\nu^* \approx 0.25$. Choosing $\nu = 0.25$ guarantees that the total cost will be smaller than that of median-of-3 and allows for a very efficient implementation of the selection of pivots, since for that choice we can avoid floating point arithmetic and integer multiplications: integer comparisons and bit shifts suffice.

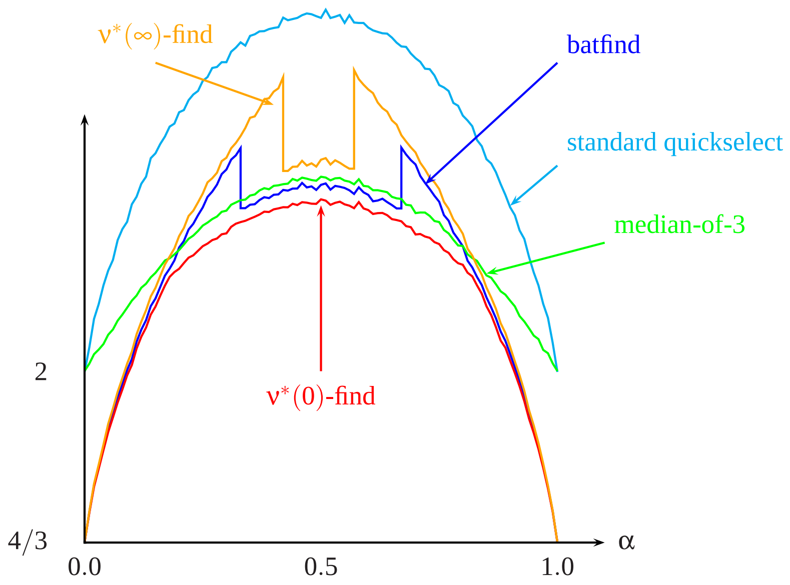


Figure 14: Plot of the experimental estimation of $f(\alpha)$ (comparisons) for several selection algorithms.

7 Experiments

We have conducted a series of experiments to compare the empirically measured performance with the analytical developments of previous sections. In these experiments, we have used arrays of $n = 10000$ elements. Five algorithms (standard quickselect, median-of-3, batfind and ν-find with $\nu = \nu^*(0) \doteq 0.182\ldots$ and $\nu = \nu^*(\infty) \doteq 0.429\ldots$) have been run for $m = 0, 100, 200, \ldots$; for each value of m, our program generates $P = 10000$ random arrays, applies all five algorithms to each of the arrays and collects the number of comparisons and exchanges made. We have include standard quickselect and median-of-three in our experiments as a further check for the experimental setup and statistical significance of the collected data.

It is important to emphasize that in our theoretical analysis of the previous sections we have considered only the asymptotic behavior and disregarded lower order terms; furthermore, the standard deviation of the investigated quantities (comparisons, exchanges, total cost) is most likely linear, like for the standard algorithm. Therefore, we can expect small but noticeable differences between the theoretical prediction and the experimental data, even for the large value of n and the large number P of tests that for each rank we have used (see Figures 14 and 15).

If we compare in each case the "theoretical prediction" $f(\alpha)$ with the measured mean number of comparisons divided by n, the relative error is usually smaller than 0.6%. There are a few ranks where the relative error (for some of the algorithms) was slightly greater than 0.6%. We also detected a maximum relative error of 9.085% in batfind at rank $j/n = 0.33$. Similar figures are obtained when analyzing the data corresponding to exchanges.

In summary, the agreement between our theoretical predictions and the experimental data is excellent, even though our analysis is asymptotic and lower order terms have not been considered.

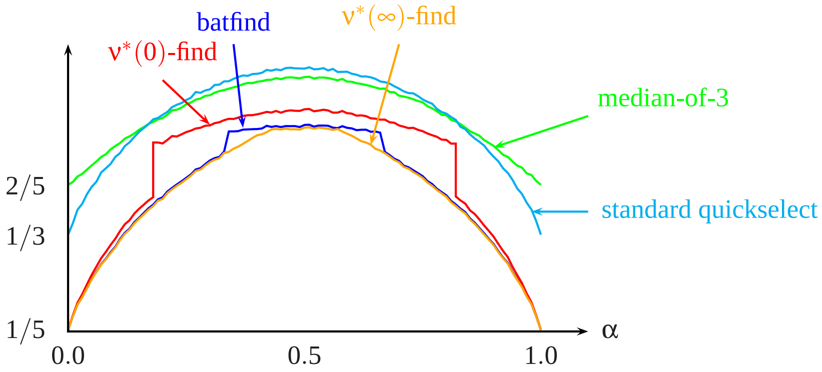


Figure 15: Plot of the experimental estimation of $t(\alpha)$ (exchanges) for several selection algorithms.

8 Optimal proportion-from-s sampling

In this section we establish the theoretical optimality of many proportion-from-s-like strategies when $s \to \infty$. It is well known [4, 5] that at least

$$n + \min(m, n-m) + o(n)$$

comparisons are necessary on the average to locate the mth smallest element out of n elements. Using the terminology of this paper, we may rephrase the main result of this section by saying that, under some additional mild circumstances, the characteristic function of proportion-from-s-like strategies is $f(\alpha) = 1 + \min(\alpha, 1-\alpha)$ when $s \to \infty$, hence optimal.

Definition 1. *An adaptive symmetric sampling strategy using samples of size s is* biased *if and only if*

$$r(\alpha) > \alpha \cdot s + 1 - \alpha, \quad \textit{for } 0 < \alpha < 1/2.$$

Notice that in a biased strategy the kth endpoint is shifted to the left of k/s when $k < s/2$ and it is shifted to the right of k/s when $k > s/2$.

Theorem 3. *For any family of* biased *sampling strategies such that* $\lim_{s\to\infty} r(\alpha)/s = \alpha$*,*

$$f^{(\infty)}(\alpha) = \lim_{s\to\infty} \lim_{n\to\infty, m/n\to\alpha} \frac{C_{n,m}}{n} = 1 + \min(\alpha, 1-\alpha).$$

Proof. The proof of theorem above amounts to showing that $f^{(\infty)}(\alpha) = 1 + \min(\alpha, 1-\alpha)$ is the unique fix point of the operator $T^{(\infty)}$, where for any $g : [0,1] \to \mathbb{R}$,

$$T^{(\infty)}(g)(\alpha) = 1 + \lim_{s\to\infty} \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \left\{ \int_\alpha^1 g\left(\frac{\alpha}{x}\right) x^{r(\alpha)} (1-x)^{s-r(\alpha)}\, dx \right.$$
$$\left. + \int_0^\alpha g\left(\frac{\alpha-x}{1-x}\right) x^{r(\alpha)-1} (1-x)^{s+1-r(\alpha)}\, dx \right\}.$$

The fact that $1+\min(\alpha,1-\alpha)$ is such a fix point is not too hard to establish once the following technical properties have been proved, for some $r \equiv r(\alpha)$ satisfying the hypotheses of the theorem:

$$\lim_{s\to\infty} \frac{s!}{(r-1)!(s-r)!} \int_\alpha^b x^{r-1}(1-x)^{s-r} y(x)\,dx = \begin{cases} y(\alpha)\,, & \text{if } \alpha < 1/2 \text{ and } \alpha < b, \quad (a) \\ 0, & \text{if } \alpha > 1/2 \text{ or } \alpha = b, \quad (b) \end{cases} \tag{11}$$

$$\lim_{s\to\infty} \frac{s!}{(r-1)!(s-r)!} \int_a^\alpha x^{r-1}(1-x)^{s-r} y(x)\,dx = \begin{cases} y(\alpha)\,, & \text{if } \alpha > 1/2 \text{ and } a < \alpha, \quad (a) \\ 0, & \text{if } \alpha < 1/2 \text{ or } a = \alpha, \quad (b) \end{cases} \tag{12}$$

$$\lim_{s\to\infty} \frac{(m+n+1)!}{m!n!} \int_a^b x^m(1-x)^n y(x)\,dx = 0, \quad \text{if } \frac{m}{m+n} \notin [a,b]. \tag{13}$$

$$\lim_{s\to\infty} \frac{(s+1)!}{r!(s-r)!} \int_{1/2}^1 x^r(1-x)^{s-r}\,dx = 1/2, \quad \text{if } \alpha = 1/2, \tag{14}$$

where $y(x)$ is an arbitrary function in $C^{(2)}[0,1]$. The proofs of these equations can be found in Appendix D.

Let $A = \int_\alpha^1 x^r(1-x)^{s-r}\,dx$, $B = \int_\alpha^1 \min(\alpha, x-\alpha) x^{r-1}(1-x)^{s-r}\,dx$, $C = \int_0^\alpha x^{r-1}(1-x)^{s+1-r}$ and $D = \int_0^\alpha \min(\alpha-x, 1-\alpha) x^{r-1}(1-x)^{s-r}$. Then, applying $T^{(\infty)}$ to $f^{(\infty)}(\alpha) = 1+\min(\alpha,1-\alpha)$ we get

$$T^{(\infty)}(f^{(\infty)})(\alpha) = 1 + \lim_{s\to\infty} \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times (A+B+C+D).$$

Now, if $\alpha < 1/2$ we have

$$\begin{aligned} C+D &= \int_0^\alpha x^{r-1}(1-x)^{s+1-r}\,dx + \int_0^\alpha (\alpha-x)x^{r-1}(1-x)^{s-r}\,dx \\ &= \int_0^\alpha (1+\alpha-2x)x^{r-1}(1-x)^{s-r}\,dx \sim 0 \end{aligned}$$

as $s \longrightarrow \infty$ because of (12b), $A \sim \alpha \cdot \frac{(r-1)!(s-r)!}{s!}$ because of (11a) with $y(x) = x$, and

$$\begin{aligned} B &= \int_{2\alpha}^1 \alpha \cdot x^{r-1}(1-x)^{s-r}\,dx + \int_\alpha^{2\alpha} (x-\alpha)x^{r-1}(1-x)^{s-r}\,dx \\ &\sim \int_\alpha^{2\alpha} (x-\alpha)x^{r-1}(1-x)^{s-r}\,dx \sim 0, \end{aligned}$$

applying both (13) and (11a) with $y(x) = (x-\alpha)$ in the second step. Altogether,

$$T^{(\infty)}(f^{(\infty)})(\alpha) \sim 1+\alpha.$$

Also, because of symmetry of $r(\alpha)$, it follows that for $\alpha > 1/2$ we have

$$T^{(\infty)}(f^{(\infty)})(\alpha) \sim 1+(1-\alpha).$$

This can also be directly proved, in the same way as we have done it for $\alpha < 1/2$.

For the special case $\alpha = 1/2$, we use the symmetry of $r(\alpha)$. We also perform the variable change $y := (1-x)$ and thus

$$A+B+C+D = 4\int_{1/2}^{1} x^r(1-x)^{s-r}\,dx - \int_{1/2}^{1} x^{r-1}(1-x)^{s-r}\,dx.$$

From there, a few straightforward manipulations, together with (14) for the first integral and the symmetry of the integrand around $x = 1/2$ for the second yield

$$T^{(\infty)}(f^{(\infty)})(\alpha) \sim 1+2\frac{r}{s+1} - \frac{1}{2} \sim 1+\frac{1}{2} = 1+\alpha = 1+(1-\alpha).$$

Hence, for all α we have just proven that $T^{(\infty)}(f^{(\infty)}) = 1+\min(\alpha, 1-\alpha) = f^{(\infty)}$. Finally, since $T^{(\infty)}$ is a contraction (see Theorem 1), it follows that $f^{(\infty)}$ must be its unique fix point. □

The previous theorem suggests that optimal performance can be achieved using variable-size samples, with s growing as n grows, as long as $s = o(n)$. If $s = \Theta(n)$ then the toll function would be $\beta n + o(n)$ for some $\beta > 1$, which precludes achieving the optimal minimum $n + \min\{m, n-m+1\} + o(n)$. We must remark that Theorem 3 concerns fixed-size sampling and considers what happens if $s \to \infty$. Hence, it does not apply to variable-size sampling. But it is rather likely that the result holds for variable-size sampling, by analogy with quicksort [18]. As long as $s = s(n)$ grows with n, the main order term would be asymptotically optimal, but the particular choice of the function s would also affect lower order terms. In order to minimize them there must be a trade-off between the quality of the pivot provided by large samples and the overhead of choosing the pivot from the sample. Based upon known results for quickselect and quicksort [18], we conjecture that the optimal size would be $s^* = \Theta(\sqrt{n})$.

These results and conjectures have undoubtedly great theoretical interest, but it is clear that quickselect with variable-sized sampling has some drawbacks for its practical application, much like Floyd and Rivest's algorithm, because of the big impact that using large samples has in the lower order terms of the performance.

9 Future work

To assess the practicality of proportion-from-s and similar variants it would be interesting to carry out a precise analysis of the lower order terms in the performance. Also a detailed analysis of the variance would be useful; we conjecture that it should be of the form $v(\alpha)\cdot n^2$, for some function v, like in the case of standard quickselect and median-of-three [11]. A careful study to establish the existence of optimal endpoints (Conjecture 1) and its behavior as a function of s would be also very interesting.

Our results of Section 8 for $s \to \infty$ suggest that variable-sized proportion-from-s sampling achieves optimal performance, but this has still to be proved. It also seems plausible that using variable-sized sampling the variance is $\Theta(\max\{n^2/s, n\cdot s\})$. It is then natural to ask ourselves about the optimal size s^* of the samples for proportion-from-s when $s = s(n) \to \infty$. As we have already discussed after Theorem 3, we conjecture that the optimal size is $s^* = \Theta(\sqrt{n})$; this choice would minimize the average number of comparisons, as well as the order of magnitude of the variance. It could also be interesting to consider strategies where the size of the samples depends on both n and α.

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 of size s is chosen as the pivot. These strategies generalize the deterministic strategies studied in this paper and include, among other, the so-called *ninther* rule or *pseudomedian-of-9* [2].

Acknowledgements

We are thankful to Sam Melkonian for his suggestions and technical advice.

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] J.L. Bentley and M.D. McIlroy. Engineering a sort function. *Software—Practice and Experience*, 23:1249–1265, 1993.

[3] N. Bleistein and R. A. Handelsman. *Asymptotic Expansions of Integrals*. Dover Pub., New York, 1975.

[4] W. Cunto and J.I. Munro. Average case selection. *Journal of the ACM*, 36(2):270–279, 1989.

[5] R.W. Floyd and R.L. Rivest. Expected time bounds for selection. *Communications of the ACM*, 18(3):165–173, 1975.

[6] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, Reading, Mass., 2nd edition, 1994.

[7] R. Grübel. On the median-of-k version of Hoare's selection algorithm. *Theoretical Informatics and Applications*, 33(2):177–192, 1999.

[8] R. Grübel and U. Rösler. Asymptotic distribution theory for Hoare's selection algorithm. *Advances in Applied Probability*, 28:252–269, 1996.

[9] C.A.R. Hoare. FIND (Algorithm 65). *Communications of the ACM*, 4:321–322, 1961.

[10] C.A.R. Hoare. Quicksort. *Computer Journal*, 5:10–15, 1962.

[11] P. Kirschenhofer and H. Prodinger. Comparisons in Hoare's Find algorithm. *Combinatorics, Probability and Computing*, 7:111–120, 1998.

[12] P. Kirschenhofer, H. Prodinger, and C. Martínez. Analysis of Hoare's FIND algorithm with median-of-three partition. *Random Structures and Algorithms*, 10(1):143–156, 1997.

[13] D.E. Knuth. Mathematical analysis of algorithms. In *Information Processing '71, Proc. of the 1971 IFIP Congress*, pages 19–27, Amsterdam, 1972. North-Holland.

[14] D.E. Knuth. *The Art of Computer Programming: Sorting and Searching*, volume 3. Addison-Wesley, Reading, Mass., 2nd edition, 1998.

[15] H.M. Mahmoud. *Sorting: A Distribution Theory*. John Wiley & Sons, New York, 2000.

[16] H.M. Mahmoud and B. Pittel. Analysis of the space of search trees under the random insertion algorithm. *Journal of Algorithms*, 10:52–75, 1989.

[17] C. Martínez, D. Panario, and A. Viola. Adaptive sampling for quickselect. In *Proc. of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'04)*, pages 440–448, 2004.

[18] C. Martínez and S. Roura. Optimal sampling strategies in quicksort and quickselect. *SIAM Journal on Computing*, 31(3):683–705, 2001.

[19] D.R. Musser. Introspective sorting and selection algorithms. *Software—Practice and Experience*, 27(8):983–993, 1997.

[20] E.D. Rainville, P.E. Bedient, and R. Bedient. *Elementary Differential Equations*. Prentice Hall, 8th edition, 1997.

[21] R. Sedgewick. *Quicksort*. Garland, New York, 1978.

[22] E.T. Whittaker and G.N. Watson. *A Course of Modern Analysis*. Cambridge University Press, 4th edition, 1927.

A Proofs of Lemmas 3 and 4

Lemma (Lemma 3). *For any adaptive sampling strategy,*

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

where $\alpha \in I_k$, $1 \le k \le \ell$.

Proof. To prove the lemma we first use the variable changes $x := \alpha/x$ and $x := (\alpha - x)/(1-x)$ in the integral equations defining the f_k's in Theorem 1 to translate them to the form

$$\begin{aligned} f_k(\alpha) = 1 + \frac{s!}{(r_k-1)!(s-r_k)!} \Bigg[& \sum_{i=0}^{s-r_k} \binom{s-r_k}{i} (-1)^{s-r_k-i} \alpha^{s+1-i} \Bigg\{ \int_{\alpha}^{a_k} \frac{f_k(x)}{x^{s+2-i}}\,dx \\ & + \sum_{d=k+1}^{\ell} \int_{I_d} \frac{f_d(x)}{x^{s+2-i}}\,dx \Bigg\} \\ & + \sum_{i=0}^{r_k-1} \binom{r_k-1}{i} (-1)^{r_k-1-i} (1-\alpha)^{s+1-i} \Bigg\{ \int_{a_{k-1}}^{\alpha} \frac{f_k(x)}{(1-x)^{s+2-i}}\,dx \\ & + \sum_{d=1}^{k-1} \int_{I_d} \frac{f_d(x)}{(1-x)^{s+2-i}}\,dx \Bigg\} \Bigg], \qquad 1 \le k \le \ell. \end{aligned} \tag{15}$$

Let

$$T(h,i,s) = \sum_{d=h+1}^{s+2} \binom{s+2}{d} \binom{d-1}{h} (i-s-2)^{\underline{d-1-h}} (s+1-i)^{\underline{s+2-d}},$$

where $x^{\underline{k}} = x \cdot (x-1) \cdots (x-k+1)$ denotes the kth falling power of x, for any $k \ge 0$ [6].

Differentiating $s+2$ times both sides of Equation (15) with respect to α, using Leibniz's rule for the Nth derivative of a product and for the derivative of integrals, yields

$$\begin{aligned} \frac{d^{s+2}}{d\alpha^{s+2}} f_k(\alpha) = \frac{s!}{(r_k-1)!(s-r_k)!} & \cdot \sum_{h=0}^{s+1} \frac{d^h}{d\alpha^h} f_k(\alpha) \\ & \cdot \Bigg(-\frac{1}{\alpha^{s+2-h}} \cdot \sum_{i=0}^{s-r_k} \binom{s-r_k}{i} (-1)^{s-r_k-i} \cdot T(h,i,s) \\ & + \frac{(-1)^{s+1-h}}{(1-\alpha)^{s+2-h}} \cdot \sum_{i=0}^{r_k-1} \binom{r_k-1}{i} (-1)^{r_k-1-i} \cdot T(h,i,s) \Bigg). \end{aligned} \tag{16}$$

It turns out that $T(h,i,s)$ has a simple closed form:

$$T(h,i,s) = (-1)^{s+1-h} \cdot i^{\underline{s+1-h}}, \tag{17}$$

and its binomial transform also has a nice closed form, namely,

$$\sum_{i=0}^{M} \binom{M}{i} (-1)^{M-i} \cdot T(h,i,s) = \begin{cases} (-1)^M \cdot M! & \text{if } h = s+1-M, \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

To prove (17) and (18) we need the following combinatorial identities that can be found, for example, in [6]:

$$\begin{aligned}
\binom{r}{k} &= \frac{r^{\underline{k}}}{k!} \\
(a+b)^n &= \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}, && \text{Newton's theorem,} \\
\binom{r}{m}\binom{m}{k} &= \binom{r}{k}\binom{r-k}{m-k}, && \text{integers } m,k \quad \text{(trinomial revision),} \\
\binom{r}{k} &= (-1)^k \binom{k-r-1}{k}, && \text{integer } k \quad \text{(upper negation),} \\
\binom{r}{k} &= \frac{r}{k}\binom{r-1}{k-1}, && \text{integer } k \neq 0 \quad \text{(absorption/extraction).}
\end{aligned}$$

Moreover, we also need Vandermonde's convolution, formula (5.24) and the upper negation of formula (5.33) in [6], namely,

$$\sum_k \binom{r}{m+k}\binom{s}{n-k} = \binom{r+s}{m+n}, \qquad \text{integers } m,n, \tag{19}$$

$$\sum_k (-1)^k \binom{l}{m+k}\binom{t+k}{n} = (-1)^{l+m}\binom{t-m}{n-l}, \quad \text{integers } l \geq 0, m, n, \tag{20}$$

$$\sum_{j\geq 0} (-1)^j \frac{\binom{n}{j}}{\binom{x+j}{j}} = \frac{x}{x+n}. \tag{21}$$

In the case of (17), we first apply the absorption formula to the second binomial coefficient $\binom{d-1}{h}$, then the trinomial revision to the product of the two binomial coefficients and finally rewrite the falling factorials as binomial coefficients to get

$$T(h,i,s) = (s+2)^{\underline{s+1-h}} \sum_{d=h+1}^{s+2} \frac{h+1}{d} \binom{i-s-2}{d-1-h}\binom{s+1-i}{s+2-d}.$$

Then, we use the upper negation formula in the first binomial coefficient and change the formal variable inside the sum to obtain

$$T(h,i,s) = (s+2)^{\underline{s+1-h}} \times \sum_{m=0}^{s+1-h} \frac{h+1}{m+h+1} (-1)^m \binom{m+1+s-i}{m}\binom{s+1-i}{s+1-m-h}.$$

Now, replace $(h+1)/(m+h+1)$ using formula (21) with $x = h+1$ and $n = m$ so that

$$T(h,i,s) = (s+2)^{\underline{s+1-h}} \times \sum_{j\geq 0} \frac{(-1)^j}{\binom{h+1+j}{j}} \cdot \sum_{m\geq 0} (-1)^m \binom{m+1+s-i}{m}\binom{m}{j}\binom{s+1-i}{s+1-m-h}.$$

We may rewrite the double sum using only factorials, multiply and divide by $(s+1-h)!$ and $(s+1-h-j)!$, express the result as another product of binomial coefficients and change the formal variable m by k to get

$$\begin{aligned} T(h,i,s) = (s+2)^{\underline{s+1-h}} \times \\ \sum_{j\geq 0}(-1)^j \frac{\binom{s+1-h}{j}}{\binom{h+1+j}{j}} \sum_{k\geq 0}(-1)^k \binom{s+1-h-j}{k-j}\binom{s+1-i+k}{s+1-h}. \end{aligned}$$

At this point we use formula (20) for $t=s+1-i$, $n=s+1-h$, $l=s+1-h-j$ and $m=-j$ to obtain

$$\begin{aligned} T(h,i,s) &= (-1)^{s+1-h}(s+2)^{\underline{s+1-h}} \sum_{j\geq 0}(-1)^j \frac{\binom{s+1-h}{j}\binom{s+1-i+j}{j}}{\binom{h+1+j}{j}} \\ &= (-1)^{s+1-h}(s+1-h)! \sum_{j\geq 0}(-1)^j \binom{s+2}{h+1+j}\binom{s+1-i+j}{j}, \end{aligned}$$

where the last identity holds after rearranging the factorials. Finally, if we apply upper negation to the last binomial coefficient and then use Vandermonde's convolution we get

$$\begin{aligned} T(h,i,s) &= (-1)^{s+1-h}(s+1-h)! \sum_{j\geq 0}\binom{s+2}{h+1+j}\binom{i-s-2}{j} \\ &= (-1)^{s+1-h}(s+1-h)! \sum_{j\geq 0}\binom{s+2}{h+1+j}\binom{i-s-2}{i-s-2-j} \\ &= (-1)^{s+1-h}(s+1-h)! \binom{i}{h+i-s-1} \\ &= (-1)^{s+1-h} i^{\underline{s+1-h}}. \end{aligned}$$

To prove (18) we first use (17) to obtain

$$\sum_{i=0}^{M}\binom{M}{i}(-1)^{M-i}\cdot T(h,i,s) = \sum_{i=0}^{M}\binom{M}{i}(-1)^{M-i}\cdot(-1)^{s+1-h}\cdot i^{\underline{s+1-h}}.$$

If we complete the binomial coefficients from the falling factorials and then we get

$$\begin{aligned} &\sum_{i=0}^{M}\binom{M}{i}(-1)^{M-i}\cdot(-1)^{s+1-h}\cdot i^{\underline{s+1-h}} \\ &\qquad = (-1)^{s+1-h}(s+1-h)!\binom{M}{s+1-h}\sum_{i=0}^{M}(-1)^{M-i}\binom{M+h-s-1}{i+h-s-1} \\ &\qquad = (-1)^{s+1-h}(s+1-h)!\binom{M}{s+1-h}\sum_{i=0}^{M}(-1)^{M-i}\binom{M+h-s-1}{M-i} \\ &\qquad = (-1)^{s+1-h}(s+1-h)!\binom{M}{s+1-h}\sum_{i=0}^{M}(-1)^{i}\binom{M+h-s-1}{i}. \end{aligned}$$

It is easy to see using Newton's theorem that the last sum is equal to zero unless $M = s+1-h$.

Plugging identities (17) and (18) into (16), a few additional manipulations yield the differential equation given in the statement of the lemma. □

Lemma (Lemma 4). *Let C_n be the average cost to select an element of random rank out of n elements using a symmetric adaptive sampling strategy. Then we have*

$$\overline{f} = \lim_{n\to\infty} \frac{C_n}{n} = \int_0^1 f(\alpha)\, d\alpha,$$

where $f(\alpha)$ is the characteristic function of the algorithm, and it is as given by Theorem 1.

Proof. Recall that, by definition,

$$C_n = \frac{1}{n} \sum_{1\le m\le n} C_{n,m}.$$

Now replace $C_{n,m}$ by its asymptotic estimate as given by Theorem 1:

$$\begin{aligned} C_n = \frac{1}{n} \sum_{1\le m\le n} C_{n,m} &= \frac{1}{n} \sum_{1\le m\le n} (f(m/n)\cdot n + o(n)) \\ &= \sum_{1\le m\le n} f(m/n) + \sum_{1\le m\le n} o(1) = \sum_{0\le m<n} f((m+1)/n) + o(n). \end{aligned}$$

If $f(\alpha)$ had no discontinuities then we could use Euler-Mclaurin formula to show that

$$C_n = \int_0^n f\left(\frac{x+1}{n}\right) dx + \sum_{k=1}^{\infty} \frac{B_k}{k!} \frac{d^{k-1} f}{dx^{k-1}}((x+1)/n)\Bigg|_0^n + o(n), \tag{22}$$

where B_k denotes the kth Bernoulli number. Using then the symmetry of $f(\alpha)$, the odd index terms cancel out each other and since the Bernoulli numbers of even index are zero, it follows that

$$C_n = \int_0^n f((x+1)/n)\, dx + o(n) = n \cdot \int_{1/n}^{1+1/n} f(y)\, dy + o(n).$$

Finally, dividing by n and passing to the limit yields the statement of the lemma.

Since in general $f(\alpha)$ has a finite number of discontinuities, but still enjoys the necessary smoothness properties piecewise, it is not difficult to adapt Euler-Mclaurin formula so that (22) can be "broken" into ℓ parts and avoid the discontinuities. Hence we get

$$\begin{aligned} C_n = n \cdot \Bigg(\int_{1/n}^{\lfloor na_1 \rfloor/n} f(y)\, dy + \int_{\lfloor na_1 \rfloor/n+1/n}^{\lfloor na_2 \rfloor/n} f(y)\, dy + \dots \\ + \int_{\lfloor na_{\ell-1} \rfloor/n+1/n}^{1+1/n} f(y)\, dy \Bigg) + o(n). \end{aligned}$$

The lemma follows dividing by n and passing to the limit. □

B Proof of Theorem 2

Theorem (Theorem 2). *There exists an optimal value of* ν*, namely* $\nu^* \doteq 0.182\ldots$*, such that* $f_{1,\nu^*}(\nu^*) = f_{2,\nu^*}(\nu^*)$ *and for all* ν*,* $0 < \nu < 1/2$*, and for all* α*,* $0 \leq \alpha \leq 1$*,*

$$f_{\nu^*}(\alpha) \leq f_\nu(\alpha).$$

Proof. Consider a function $f : [0,1] \to \mathbb{R}$ and some adaptive sampling strategy as given by $a_0, a_1, \ldots, a_\ell$ and $r(\alpha)$. Let $\hat{T}$ be the functional operator

$$\begin{aligned}\hat{T}_k(f)(\alpha) = \frac{s!}{(r_k-1)!(s-r_k)!}\Bigg[&\int_{\alpha/a_k}^{1} f_k(\alpha/x)\, x^{r_k}(1-x)^{s-r_k}\, dx\\ &+ \int_0^{\frac{\alpha-a_{k-1}}{1-a_{k-1}}} f_k\left(\frac{\alpha-x}{1-x}\right) x^{r_k-1}(1-x)^{s+1-r_k}\, dx\\ &+ \sum_{d=k+1}^{\ell} \int_{I'_d} f_d(\alpha/x)\, x^{r_k}(1-x)^{s-r_k}\, dx\\ &+ \sum_{d=1}^{k-1} \int_{I''_d} f_d\left(\frac{\alpha-x}{1-x}\right) x^{r_k-1}(1-x)^{s+1-r_k}\, dx\Bigg],\end{aligned}$$

with $I'_d = (\alpha/a_d, \alpha/a_{d-1})$ and $I''_d = \left(\frac{\alpha-a_d}{1-a_d}, \frac{\alpha-a_{d-1}}{1-a_{d-1}}\right)$, and f_k the restriction of f to the kth interval.

It turns out that this operator is a contraction; this can be proved in a similar way as we proved in Theorem 1 that $T = 1 + \hat{T}$ is a contraction. Furthermore, $\hat{T}$ is linear, which proves the following

$$f_k = \hat{T}_k(f) \text{ for all } 1 \leq k \leq \ell \text{ implies } f = 0. \tag{23}$$

Let $g_1(\nu) = f_{1,\nu}(\nu)$ and $g_2(\nu) = f_{2,\nu}(\nu)$. The respective limits when $\nu \to 0$ are $g_1(0) = 3/2$ and $g_2(0) = 2$. On the other hand, when $\nu \to 1/2$ we have $g_1(1/2) > 3$ and $g_2(1/2) < 3$. Since both functions are strictly increasing in $(0,1/2)$—their derivatives w.r.t. ν are strictly positive—, it follows that $g_1(\nu) = g_2(\nu)$ has a unique solution, say ν^*, in the interval $(0,1/2)$.

Take $f_{1,\nu} = 1 + \hat{T}_1(f_{1,\nu}, f_{2,\nu})$ and $f_{2,\nu} = 1 + \hat{T}_2(f_{1,\nu}, f_{2,\nu})$. Differentiating both equations with respect to ν and setting $\nu = \nu^*$ many terms cancel out because $f_{1,\nu^*}(\nu^*) = f_{2,\nu^*}(\nu^*)$, so we finally arrive at

$$\begin{aligned}\left.\frac{\partial f_{1,\nu}}{\partial \nu}\right|_{\nu=\nu^*} &= \hat{T}_1\left(\left.\frac{\partial f_{1,\nu}}{\partial \nu}\right|_{\nu=\nu^*}, \left.\frac{\partial f_{2,\nu}}{\partial \nu}\right|_{\nu=\nu^*}\right),\\ \left.\frac{\partial f_{2,\nu}}{\partial \nu}\right|_{\nu=\nu^*} &= \hat{T}_2\left(\left.\frac{\partial f_{1,\nu}}{\partial \nu}\right|_{\nu=\nu^*}, \left.\frac{\partial f_{2,\nu}}{\partial \nu}\right|_{\nu=\nu^*}\right).\end{aligned}$$

Hence, by (23), it follows that $\left.\frac{\partial f_{1,\nu}}{\partial \nu}\right|_{\nu=\nu^*}(\alpha) = \left.\frac{\partial f_{2,\nu}}{\partial \nu}\right|_{\nu=\nu^*}(\alpha) = 0$ for any α in the corresponding intervals.

Also, if we compute the second derivatives of $f_{1,\nu}(\alpha)$ and $f_{2,\nu}(\alpha)$ w.r.t. ν and set $\nu = \nu^*$, both are strictly positive in the appropriate intervals. Indeed,

$$\frac{\partial^2 f_{1,\nu}}{\partial \nu^2}\bigg|_{\nu=\nu^*} = \hat{T}_1\left(\frac{\partial^2 f_{1,\nu}}{\partial \nu^2}\bigg|_{\nu=\nu^*} \frac{\partial^2 f_{2,\nu}}{\partial \nu^2}\bigg|_{\nu=\nu^*}\right) + \Delta_1(\alpha),$$
$$\frac{\partial^2 f_{2,\nu}}{\partial \nu^2}\bigg|_{\nu=\nu^*} = \hat{T}_2\left(\frac{\partial^2 f_{1,\nu}}{\partial \nu^2}\bigg|_{\nu=\nu^*}, \frac{\partial^2 f_{2,\nu}}{\partial \nu^2}\bigg|_{\nu=\nu^*}\right) + \Delta_2(\alpha),$$

where $\Delta_1(\alpha)$ is strictly positive in the interval $(0,\nu^*]$ and $\Delta_2(\alpha)$ is strictly positive in the interval $(\nu^*, 1/2)$, and this property translates to the second derivatives of $f_{1,\nu}(\alpha)$ and $f_{2,\nu}(\alpha)$. Hence, $\nu = \nu^*$ is a local minimum.

The limit values $\nu \to 0$ (median-of-3) and $\nu \to 1/2$ are not minimum, hence to complete the proof we need to show that there are no additional local extrema of $f_\nu(\alpha)$ for fixed α. This can be shown by contradiction. If we assume that there exists some $\nu^{**} \neq \nu^*$ such that $\frac{\partial f_\nu}{\partial \nu}\Big|_{\nu=\nu^{**}}(\alpha) = 0$ then this implies that $f_{1,\nu^{**}}(\nu^{**}) = f_{2,\nu^{**}}(\nu^{**})$; but we already know that there is only a unique value of ν where this happens. □

Also, it is worth mentioning that Conjecture 1 in Section 5 could be proven using the same strategy as above. We should establish the existence of a unique $\mathbf{a}^*$ such that $f^{(s)}_{\mathbf{a}^*}(\alpha)$ is continuous. If we compute $\partial f^{(s)}_{\mathbf{a}}(\alpha)/\partial a_i$ for $1 \le i \le s$, it is fairly easy to show that all these derivatives vanish at $\mathbf{a} = \mathbf{a}^*$ for any value of α, because $f^{(s)}_{\mathbf{a}^*} = \hat{\mathbf{T}}(f^{(s)}_{\mathbf{a}^*})$. Conversely, if $\partial f^{(s)}_{\mathbf{a}^*}(\alpha)/\partial a_i = 0$ for some $\mathbf{a}^*$ and all α and all $1 \le i \le s$ then it is not difficult to prove that $f^{(s)}_{\mathbf{a}^*}(\alpha)$ is continuous for $\alpha \in [0,1]$. The proof could then be completed by proving that the particular quadratic form corresponding to $f^{(s)}_{\mathbf{a}^*}$ is positive definite.

C Coefficients of ν-find

The reader will notice that the expressions for C_2, C_5 and C_7 given below are in terms of $\Delta := 70\nu^5 - 210\nu^4 + 294\nu^3 - 224\nu^2 + 90\nu - 15$ and the integrals $\mathbf{A}_i(\nu) = \int_0^\nu \kappa_\nu(u)/u^i\,du$ and $\mathbf{B}_i(\nu) = \int_0^\nu \kappa_\nu(u)/(1-u)^i\,du$, where $\kappa_\nu(u) = C_3(\nu)\cdot K_1(u) + C_4(\nu)\cdot K_2(u)$. Hence, the constants C_2, C_5 and C_7 are given in terms of the (unknown) values $C_3(\nu)$ and $C_4(\nu)$. The last two equations in the list below, with the integrals in the left hand side, allow us to recover the values of C_3 and C_4, and from there the remaining C_i's. Actually, it is not very difficult to find closed forms for the C_i's, but the resulting expressions are much lengthier and cumbersome to handle.

We give the C_i's that correspond to the average total cost (see Section 6). Setting $\xi_1 = 1, \xi_2 = 0$, we can obtain the constants for the average number of comparisons, and with $\xi_1 = 0, \xi_2 = 1$, we obtain those for the average number of exchanges. Also, setting $\nu = 1/3$, we can get the values corresponding to pure proportion-from-3 (Section 4).

- $C_0 = -\frac{24}{11}(\xi_1 + 3\xi_2/20)$, $\quad C_1 = -\frac{28}{33}(\xi_1 + 3\xi_2/20)$, $\quad C_6 = \frac{7}{4}C_5 + 3\cdot(\xi_1 + \xi_2/5)$,

- $\Delta \cdot C_7 =$

$$\begin{aligned}
&\xi_1 \cdot \Bigg(-\frac{8}{11}(55\nu^5 - 180\nu^4 + 309\nu^3 - 334\nu^2 + 204\nu - 45)\ln(1-\nu) \\
&\qquad - \frac{1}{33}(350\nu^8 - 3360\nu^7 + 12180\nu^6 - 21680\nu^5 + 23655\nu^4 - 18480\nu^3 \\
&\qquad\quad + 12128\nu^2 - 5436\nu + 990)\Bigg) \\
&+\xi_2 \cdot \Bigg(\frac{6}{55}(15\nu^4 - 78\nu^3 + 158\nu^2 - 138\nu + 45)\ln(1-\nu) \\
&\qquad - \frac{1}{110}(175^8 - 1680\nu^7 + 6090\nu^6 - 11060\nu^5 + 12240\nu^4 - 9636\nu^3 \\
&\qquad\quad + 6174\nu^2 - 2916\nu + 660)\Bigg) \\
&+\Big(36(2\nu-3)\nu^4\ln(1-\nu) - 3(35\nu^4 - 84\nu^3 + 84\nu^2 + 20\nu - 36)\nu^4\Big) \cdot \mathbf{A}_4(\nu) \\
&-\Big(36(2\nu^2 - 3\nu + 9)(\nu-1)^3\ln(1-\nu) \\
&\qquad + 3(35\nu^6 - 266\nu^5 + 651\nu^4 - 832\nu^3 + 706\nu^2 - 492\nu + 168)(\nu-1)^2\Big) \cdot \mathbf{B}_4(\nu) \\
&-\Big(432(\nu-1)^4\ln(1-\nu) \\
&\qquad - 12(35\nu^5 - 105\nu^4 + 126\nu^3 - 112\nu^2 + 117\nu - 51)(\nu-1)^3\Big) \cdot \mathbf{B}_5(\nu),
\end{aligned}$$

- $\Delta \cdot C_2 =$

$$\begin{aligned}
&\xi_1 \cdot \Bigg(\frac{8}{11}(55\nu^5 - 180\nu^4 + 309\nu^3 - 334\nu^2 + 204\nu - 45)\ln\frac{\nu}{1-\nu} \\
&\qquad - \frac{1}{33}\frac{2370\nu^5 - 7020\nu^4 + 7796\nu^3 - 2568\nu^2 - 375\nu + 220}{\nu}\Bigg) \\
&-\xi_2 \cdot \Bigg(\frac{6}{55}(15\nu^4 - 78\nu^3 + 158\nu^2 - 138\nu + 45)\ln\frac{\nu}{1-\nu} \\
&\qquad + \frac{1}{110}\frac{525\nu^5 - 1860\nu^4 + 1896\nu^3 - 228\nu^2 - 600\nu + 220}{\nu}\Bigg) \\
&-\Bigg(36(2\nu-3)\nu^4\ln\frac{\nu}{1-\nu} + 3(81\nu^2 - 60\nu + 10)\nu^2\Bigg) \cdot \mathbf{A}_4(\nu) \\
&+\Bigg(36(2\nu^2 - 3\nu + 9)(\nu-1)^3\ln\frac{\nu}{1-\nu} \\
&\qquad -3\frac{(129\nu^3 - 174\nu^2 - 25\nu + 40)(\nu-1)^2}{\nu}\Bigg) \cdot \mathbf{B}_4(\nu) \\
&+\Bigg(432(\nu-1)^4\ln\frac{\nu}{1-\nu} - 60\frac{(2\nu-1)(7\nu^2 - 7\nu - 2)(\nu-1)^3}{\nu}\Bigg) \cdot \mathbf{B}_5(\nu),
\end{aligned}$$

- $\Delta \cdot C_5 =$

$$
\begin{aligned}
&-\frac{8}{11}\xi_1(155\nu^4-450\nu^3+573\nu^2-338\nu+66)\nu\\
&-\frac{6}{55}\xi_2(210\nu^4-615\nu^3+804\nu^2-514\nu+132)\nu\\
&-36(2\nu-3)\nu^4\mathbf{A}_4(\nu)+36(2\nu^2-3\nu+9)(\nu-1)^3\mathbf{B}_4(\nu)+432(\nu-1)^4\mathbf{B}_5(\nu),
\end{aligned}
$$

- $\Delta \cdot (\mathbf{A}_3(\nu)+\mathbf{B}_3(\nu)) =$

$$
\begin{aligned}
&\xi_1\cdot\left(-\frac{1}{11}(140\nu^5-630\nu^4+1680\nu^3-2660\nu^2+2046\nu-495)\ln\frac{\nu}{1-\nu}\right.\\
&\qquad\left.-\frac{2660\nu^8-14630\nu^7+35882\nu^6-43624\nu^5+20042\nu^4+5729\nu^3-7947\nu^2+2070\nu-110}{66\nu^2(\nu-1)}\right)\\
&-\xi_2\cdot\left(\frac{3}{110}(70\nu^5-315\nu^4+840\nu^3-1330\nu^2+1056\nu-330)\ln\frac{\nu}{1-\nu}\right.\\
&\qquad\left.+\frac{1330\nu^8-7315\nu^7+17941\nu^6-21812\nu^5+10516\nu^4+2694\nu^3-4782\nu^2+1530\nu-110}{220\nu^2(\nu-1)}\right)\\
&+\left(63(2\nu-3)\nu^4\ln\frac{\nu}{1-\nu}-\frac{3}{2}\frac{(70\nu^5-511\nu^4+833\nu^3-518\nu^2+150\nu-20)\nu}{\nu-1}\right)\cdot\mathbf{A}_4(\nu)\\
&-\left(63(2\nu^2-3\nu+9)(\nu-1)^3\ln\frac{\nu}{1-\nu}\right.\\
&\qquad\left.+\frac{3}{2}\frac{(70\nu^6-539\nu^5+1029\nu^4-315\nu^3-453\nu^2+248\nu-20)(\nu-1)}{\nu^2}\right)\cdot\mathbf{B}_4(\nu)\\
&-\left(756(\nu-1)^4\ln\frac{\nu}{1-\nu}-6\frac{(2\nu-1)(105\nu^4-210\nu^3+49\nu^2+56\nu-5)(\nu-1)^2}{\nu^2}\right)\cdot\mathbf{B}_5(\nu),
\end{aligned}
$$

- $\Delta \cdot \mathbf{A}_5(\nu) =$

$$
\begin{aligned}
&\xi_1\cdot\frac{5}{198}\frac{210\nu^8-1092\nu^7+2562\nu^6-3546\nu^5+3195\nu^4-1880\nu^3+687\nu^2-138\nu+11}{\nu^4}\\
&+\xi_2\cdot\frac{1}{132}\frac{105\nu^8-546\nu^7+1281\nu^6-1806\nu^5+1680\nu^4-1050\nu^3+426\nu^2-102\nu+11}{\nu^4}\\
&+\frac{1}{2}\frac{105\nu^5-294\nu^4+399\nu^3-300\nu^2+120\nu-20}{\nu}\cdot\mathbf{A}_4(\nu)\\
&+\frac{1}{2}\frac{(105\nu^5-231\nu^4+273\nu^3-183\nu^2+66\nu-10)(\nu-1)^3}{\nu^4}\cdot\mathbf{B}_4(\nu)\\
&+\frac{(70\nu^5-140\nu^4+154\nu^3-98\nu^2+34\nu-5)(\nu-1)^4}{\nu^4}\cdot\mathbf{B}_5(\nu).
\end{aligned}
$$

D Proof of Equations (11)-(14)

In this appendix we give the proof of the equations (11) to (14) given in Section 8:

$$\lim_{s\to\infty}\frac{s!}{(r-1)!(s-r)!}\int_{\alpha}^{b}x^{r-1}(1-x)^{s-r}y(x)\,dx=\begin{cases}y(\alpha), & \text{if } \alpha<1/2 \text{ and } \alpha<b, \quad (a)\\ 0, & \text{if } \alpha>1/2 \text{ or } \alpha=b, \quad (b)\end{cases} \tag{11}$$

$$\lim_{s\to\infty}\frac{s!}{(r-1)!(s-r)!}\int_{a}^{\alpha}x^{r-1}(1-x)^{s-r}y(x)\,dx=\begin{cases}y(\alpha), & \text{if } \alpha>1/2 \text{ and } a<\alpha, \quad (a)\\ 0, & \text{if } \alpha<1/2 \text{ or } a=\alpha, \quad (b)\end{cases} \tag{12}$$

$$\lim_{s\to\infty}\frac{(m+n+1)!}{m!n!}\int_{a}^{b}x^{m}(1-x)^{n}y(x)\,dx=0, \quad \text{if } \frac{m}{m+n}\notin[a,b]. \tag{13}$$

$$\lim_{s\to\infty}\frac{(s+1)!}{r!(s-r)!}\int_{1/2}^{1}x^{r}(1-x)^{s-r}\,dx=1/2, \quad \text{if } \alpha=1/2, \tag{14}$$

where $y(x)$ is an arbitrary function in $C^{(2)}[0,1]$.

Our starting point is the Eulerian integral of the first kind [22][4]:

$$\int_{0}^{1}x^{u}(1-x)^{v}\,dx=\frac{u!v!}{(u+v+1)!}. \tag{24}$$

Intuitively, the argument to prove Equations (11) to (13) is the following. When s is large, the integrand, say $x^{r-1}(1-x)^{s-r}y(x)$, is highly concentrated around $x=(r-1)/(s-1)$, and hence if the interval of integration does not contain $(r-1)/(s-1)$ then the integral is 0 (as stated by (11b), (12b) and (13)). On the other hand, if we integrate an interval that properly contains $(r-1)/(s-1)$ then we can safely extend the interval of integration to $[0,1]$ and apply (24), giving cases (11a) and (12a). If $\alpha=1/2$ then $r/s\sim 1/2$ (because r is symmetric) and half of the weight of the integrand goes to each side of $x=1/2$, hence (14).

The rigorous proof of all the integrals is based on Laplace's method. We prove here only one of the equations, namely, Equation (11), the other proofs are quite similar.

Let

$$I(\omega)=\int_{a}^{b}[\phi(x)]^{\omega}y(x)\,dx,$$

where ϕ is in $C^{(4)}[a,b]$ and nonnegative, and $y\in C^{(2)}[a,b]$. If the absolute maximum of $\phi(x)$ on $[a,b]$ occurs at $x=c$ with $a<c<b$, $\phi'(c)=0$ and $\phi''(c)<0$, then (see for instance [3, Ch. 5])

$$I(\omega)=\frac{(\phi(c))^{\omega+1/2}}{\omega^{1/2}}y(c)\sqrt{\frac{2\pi}{|\phi''(c)|}}\left(1+O\left(\frac{1}{\omega^{1/2}}\right)\right), \tag{25}$$

[4] We use $z!$ instead of $\Gamma(z+1)$.

whereas if the absolute maximum occurs at $x = a$ then[5]

$$I(\omega) = \frac{(\phi(a))^{\omega+1/2}}{\omega^{1/2}} y(a) \sqrt{\frac{\pi}{2|\phi''(a)|}} \left(1 + O\left(\frac{1}{\omega^{1/2}}\right)\right). \tag{26}$$

Consider (11a). The absolute maximum of $\phi(x) = x^{(r-1)/s}(1-x)^{(s-r)/s}$ in $[0,1]$ is at $x = (r-1)/(s-1)$. Since r is biased (see Definition 1 in Section 8) and $r/s \to \alpha < 1/2$, we have $\alpha < (r-1)/(s-1) < b$ for all sufficiently large s. Hence, we apply (25) with $\omega = s$ and $c = (r-1)/(s-1)$ to get

$$\begin{aligned} I(s) = \int_\alpha^b & x^{r-1}(1-x)^{s-r} y(x) \\ & \sim y\left(\frac{r-1}{s-1}\right) \sqrt{2\pi} \sqrt{\frac{(r-1)(s-r)}{(s-1)^3}} \left(\frac{r-1}{s-1}\right)^{r-1} \left(\frac{s-r}{s-1}\right)^{s-r}. \end{aligned}$$

Then we multiply by $s!/(r-1)!(s-r)!$ and take the limit when $s \to \infty$. Applying Stirling's asymptotic estimate for $z!$ we obtain the stated result:

$$\lim_{s\to\infty} \frac{s!}{(r-1)!(s-r)!} I(s) = \lim_{s\to\infty} y\left(\frac{r-1}{s-1}\right) \left(\frac{s}{s-1}\right)^{s-1} e^{-1} \sqrt{\left(\frac{s}{s-1}\right)^3} = y(\alpha).$$

On the other hand for (11b), since r is biased by hypothesis we have that $(r-1)/(s-r)$ is outside $[\alpha, b]$ and then the absolute maximum of $\phi(x) = x^{(r-1)/s}(1-x)^{(s-r)/s}$ is located at $x = \alpha$. Since we assume now that $\alpha > 1/2$, the hypotheses of Theorem 3 imply that $\delta = \alpha \cdot s - r(\alpha)$ is positive and $\delta = o(s)$; if $\delta \to \infty$ then the result is easier to prove, so we assume further $\delta = \Theta(1)$. Then we have

$$I(s) \sim y(\alpha)\alpha^r(1-\alpha)^{s+1-r} \sqrt{\frac{\pi s(s-1)}{2((r-1)(s-r)s - (r-1)\delta^2 - (s-r)\delta^2 + s\delta^2)}}.$$

Again, multiplying by $s!/(r-1)!(s-r)!$ and taking the limit when $s \to \infty$, we obtain:

$$\begin{aligned} & \frac{s!}{(r-1)!(s-r)!} I(s) \\ & \quad \sim y(\alpha)\frac{1}{2}(r-1+\delta)(s-r-\delta)\frac{1}{(s-1)^2} \sqrt{\frac{s^2(s-1)}{(r-1)(s-r)((r-1)(s-r)s+\delta^2)}}, \end{aligned}$$

and using $s-1 \sim s$ we finally get

$$\frac{s!}{(r-1)!(s-r)!} I(s) \sim \frac{y(\alpha)}{2s};$$

hence its limit is 0 when $s \to \infty$.

[5] The same formula applies if the absolute maximum occurs at $x = b$, replacing all a's by b's.