On the Variance of Quickselect*

Jean Daligault† Conrado Martínez‡

December 20, 2005

Abstract

Quickselect with median-of-three is routinely used as the method of choice for selection of the mth element out of n in general-purpose libraries such as the C++ *Standard Template Library*. Its average behavior is fairly well understood and has been shown to outperform that of the standard variant, which chooses a random pivot on each stage. However, no results were previously known about the variance of the median-of-three variant, other than for the number of comparisons made when the rank m of the sought element is given by a uniform random variable. Here, we consider the variance of the number of comparisons made by quickselect with median-of-three and other quickselect variants when selecting the mth element for $m/n \to \alpha$ as $n \to \infty$. We also investigate the behavior of proportion-from-s sampling as $s \to \infty$.

1 Introduction

Hoare's quickselect [5] 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 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, now looking for the $(m - j)$th element.

The main measure of quickselect's performance is the number $C^{(0)}_{n,m}$ of comparisons made to select the mth smallest element out of n. Knuth [8] has shown that

$$\mathbb{E}\left[C^{(0)}_{n,m}\right] = 2\,(n + 3 + (n+1)H_n - (m+2)H_m - (n+3-m)H_{n+1-m})\,,$$

where $H_n = \sum_{1\le j\le n} 1/j$ is the nth harmonic number. Thus $\mathbb{E}\left[C^{(0)}_{n,m}\right]$ is $\Theta(n)$ for all values of m, $1 \le m \le n$. Another interesting measure of performance is the number $C^{(0)}_n$ of comparisons needed when the rank of the sought element is given by a uniformly distributed random variable in $\{1, \ldots, n\}$. Since $\mathbb{E}\left[C^{(0)}_n\right] = (1/n) \cdot \sum_{1\le m\le n} \mathbb{E}\left[C^{(0)}_{n,m}\right]$ it immediately follows that $\mathbb{E}\left[C^{(0)}_n\right] = 3n + o(n)$.

Sometimes it is more convenient (and more amenable to analysis) to consider the asymptotic behavior of $C^{(0)}_{n,m}$ as $n \to \infty$ and $m/n \to \alpha$, for some fixed α, $0 \le \alpha \le 1$. It is not difficult to show that $m_0(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \mathbb{E}\left[C^{(0)}_{n,m}\right] / n = 2 + 2\mathcal{H}(\alpha)$, where $\mathcal{H}(x) = -(x \log x + (1-x)\log(1-x))$ is the *entropy* function.

Kirschenhofer and Prodinger [6] have computed the exact form of $\mathbb{V}\left[C^{(0)}_{n,m}\right]$. It is $\Theta(n^2)$, but even its asymptotic behavior for $m = \alpha \cdot n + o(n)$ is expressed by a rather complicated formula:

(1.1)

$$\begin{aligned} v_0(\alpha) &= \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \mathbb{V}\left[C^{(0)}_{n,m}\right] / n^2 \\ &= -2\mathcal{H}^2(\alpha) + 4\mathcal{H}(\alpha) - 4\log(\alpha)\log(1-\alpha) \\ &\quad + (5 + \frac{2\pi^2}{3})\alpha(1-\alpha) \\ &\quad + \frac{1}{2} - 4\alpha\,\mathrm{dilog}\,\alpha - 4(1-\alpha)\,\mathrm{dilog}(1-\alpha), \end{aligned}$$

where $\mathrm{dilog}\, x = \int_1^x \frac{\log z}{1-z}\, dz$ denotes the dilogarithm [1].

In *quickselect with median-of-three* the pivot of each recursive stage is the median of a sample of three elements of the array. This reduces the probability of uneven partitions and there is a correspond-

*The research of the authors was supported by the Spanish Min. of Science and Technology project TIC2002-00190 (AEDRI II).

†ENS Cachan. 94235 Cachan Cedex, France. jean_daligault at yahoo dot fr.

‡Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. E-08034 Barcelona, Spain. conrado at lsi dot upc dot edu.

ing reduction in the average performance (see [4, 7] and references therein). In particular, $m_1(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \mathbb{E}\left[C_{n,m}^{(1)}\right] / n = 2 + 3\alpha(1-\alpha)$, and the average number of comparisons to locate an element of random rank is $\mathbb{E}\left[C_n^{(1)}\right] = 5/2\, n + o(n)$. The generalization to *quickselect with median-of-*$(2t+1)$ has also been considered, both for fixed t and for variable-sized samples, i.e., when $t = t(n)$. Grübel [4] has investigated the properties of $m_t(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \mathbb{E}\left[C_{n,m}^{(t)}\right] / n$. Martínez and Roura [10] have computed the expected value and variance of the number of comparisons needed to locate an element of random rank $C_n^{(t)}$, for all fixed t. They also establish results for variable-size samples $(t = t(n))$, namely, the optimal sample size.

Martínez, Panario and Viola [9] have considered another family of sampling strategies that they call *proportion-from-s*. At each recursive stage, the chosen pivot has a relative rank within the sample as close as possible to the current relative rank $\alpha = m/n$ of sought element. As the algorithm proceeds, the size n of current subarray and the rank m of the sought element within the current subarray change and so does the relative rank; thus the rank of the selected pivot nicely "adapts" to the current input. Their results were established in a quite general framework, which encompasses the proportion-from-s sampling strategies, the standard variant and the median-of-$(2t+1)$ sampling strategies as particular instances of so-called *adaptive sampling strategies*[1].

The goal of this paper is to investigate and obtain explicit results about

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

for several distinct variants of quickselect, each using its own sampling strategy.

First, in Section 2 we establish that, for any adaptive sampling strategy, $g(\alpha) = \lim_{\substack{n\to\infty \\ m/n\to\alpha}} \mathbb{E}\left[C_{n,m}^{\underline{2}}\right] / n^2$ exists, with $x^{\underline{k}} = x(x-1)(x-2)\cdots(x-k+1)$ denoting the kth falling factorial of x [3], and we also give there the integral equation that $g(\alpha)$ satisfies. We get then a few general results about $g(\alpha)$. The techniques used are those developed in [4, 9]. In Section 3 we obtain explicit solutions to that integral equation for the particular case of median-of-three and thus for its variance. Afterwards, in Section 4, we show that if $s \to \infty$ then proportion-from-s sampling strategies achieve not only optimal expected performance (a result due to [9]) but subquadratic variance, i.e., $\lim_{\substack{n\to\infty \\ m/n\to\alpha}} \mathbb{V}[C_{n,m}] / n^2 = 0$. The immediate consequence is that $C_{n,m}$ exhibits concentration in probability. Median-of-$(2t+1)$ also achieves subquadratic variance in the limit $t \to \infty$, even though the expected performance is not optimal in that case.

[1] Even though some of them do not "adapt" their choice of pivots, like standard quickselect; but they are special degenerate cases of the general definition given there.

2 General results

Following [9], we say that a sampling strategy is *adaptive* if it can be fully described by a function $r : [0,1] \to \{1, \ldots, s\}$, where s is the size of the samples. Quickselect with adaptive sampling works as follows: if $n \geq s$, a random sample of s elements from the current subarray of size n is chosen and the element whose rank within the sample is $r = r(\alpha)$ is picked as the pivot of the current recursive stage, where $\alpha = m/n$ is the current relative rank of the sought element. We assume further that the function r can be finitely specified by the image of each interval of a partition of $[0,1]$ into ℓ intervals. For convenience, we will assume that the intervals I_k are defined by $\ell - 1$ endpoints $0 = a_0 < a_1 < a_2 < \cdots < a_{\ell-1} < a_\ell = 1$ as follows: $I_1 = [0, a_1]$, $I_\ell = [a_{\ell-1}, 1]$, $I_k = (a_{k-1}, a_k]$ if $k > 1$ and $a_k \leq 1/2$, $I_k = [a_{k-1}, a_k)$ if $k < \ell$ and $a_{k-1} > 1/2$, and $I_k = (a_{k-1}, a_k)$ if $a_{k-1} \leq 1/2 < a_k$ and $1 < k < \ell$. We will use the notation r_k for the value of $r(\alpha)$ when $\alpha \in I_k$.

When $s = 1$ we have standard quickselect, and $r(\alpha) = r_1 = 1$ for all $\alpha \in [0,1]$. Median-of-$(2t+1)$ sampling is characterized by $s = 2t+1$ and $r(\alpha) = r_1 = t+1$ for all $\alpha \in [0,1]$. In [9], proportion-from-s sampling is introduced; for these strategies $\ell = s$ and $r(\alpha) = r_k = k$ for all $\alpha \in I_k$ and $1 \leq k \leq s$. The choice of endpoints gives raise to interesting mathematical phenomena with relevant practical implications; for "pure" proportion-from-s, we have $a_k = k/s$.

We now restate two of the fundamental results of [9] concerning quickselect with adaptive sampling.

THEOREM 2.1. ([9]) *Let $C_{n,m}$ be the 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* expectation characteristic function *of the algorithm*

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

is well defined, and

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

It is important to notice that because $r(\alpha)$ is an integer function the discontinuities carry on to $f(\alpha)$. So in general $f(\alpha)$ is defined by ℓ pieces, say, $f_1, \ldots, f_\ell$, with f_k the restriction of $f(\alpha)$ for α in the kth interval. The integral equation above can be transformed, after careful manipulations, to a set of higher-order linear differential equations, as shown in the following lemma.

LEMMA 2.1. ([9]) *For any adaptive sampling strategy,*

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

where $f(\alpha)$ is the strategy's expectation characteristic function, and $\alpha \in I_k$, $1 \le k \le \ell$.

In order to obtain similar results about the variance of $C_{n,m}$, we consider its second factorial moment $\mathbb{E}\left[C_{n,m}^{\underline{2}}\right] = \mathbb{E}[C_{n,m}(C_{n,m}-1)]$ since $\mathbb{V}[C_{n,m}] = \mathbb{E}\left[C_{n,m}^{\underline{2}}\right] + \mathbb{E}[C_{n,m}] - \mathbb{E}[C_{n,m}]^2$. The starting point of our analysis is the recurrence satisfied by $C_{n,m}(v)$, the probability generating function (PGF) of $C_{n,m}$

$$\begin{aligned} C_{n,m}(v) &= \sum_{k\ge 0} \Pr\{C_{n,m}=k\} v^k \\ &= v^{n-1}\Big[\sum_{j=1}^{m-1} \pi_{n,j}^{(s,r)} C_{n-j,m-j}(v) + \pi_{n,m}^{(s,r)} + \sum_{j=m+1}^{n} \pi_{n,j}^{(s,r)} C_{j-1,m}(v)\Big], \end{aligned} \tag{2.2}$$

where $\pi_{n,j}^{(s,r)}$ denotes the probability that the rth element of the sample of size s is the jth element among the n elements. The recurrence accounts for the $n-1$ comparisons needed to partition the array around the pivot, but it disregards the comparisons needed to select the pivot from the sample; this is unimportant since we assume that s is fixed. Now, $\mathbb{E}[C_{n,m}] = C'_{n,m}(1)$ and $\mathbb{E}\left[C_{n,m}^{\underline{2}}\right] = C''_{n,m}(1)$; hence,

$$\begin{aligned} \mathbb{E}\left[C_{n,m}^{\underline{2}}\right] &= 2(n-1)\,\mathbb{E}[C_{n,m}] - n(n-1) \\ &\quad + \sum_{j=1}^{m-1} \pi_{n,j}^{(s,r)}\, \mathbb{E}\left[C_{n-j,m-j}^{\underline{2}}\right] \\ &\quad + \sum_{j=m+1}^{n} \pi_{n,j}^{(s,r)}\, \mathbb{E}\left[C_{j-1,m}^{\underline{2}}\right]. \end{aligned}$$

From this recurrence, we can establish the following result.

THEOREM 2.2. *Let $C_{n,m}$ be cost to select the mth out of n elements using an adaptive sampling strategy with $m/n \to \alpha$ for $0 \le \alpha \le 1$ as $n \to \infty$. Then we have that the* second factorial moment characteristic function *of the algorithm*

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

is well defined, and

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

where $f(\alpha)$ is the expectation characteristic function of the sampling strategy.

Once we have results about $g(\alpha)$, they can be easily translated to the variance since $v(\alpha) = \lim_{n\to\infty} \mathbb{V}[C_{n,m}]/n^2 = g(\alpha) - f^2(\alpha)$. For instance, a sampling strategy is said to be *symmetric* if $\lim_{z\to\alpha^+} r(z) = \lim_{z\to\alpha^+} s + 1 - r(1-z)$. This notion is quite natural, and both median-of-$(2t+1)$ and proportion-from-s strategies[2] are symmetric. If r is symmetric then both $g(\alpha)$ and $v(\alpha)$ are as well, now in the usual sense, i.e., $g(\alpha) = g(1-\alpha)$ and $v(\alpha) = v(1-\alpha)$. Another interesting result concerns the behavior of $v(\alpha)$ when $\alpha \to 0$.

[2]Provided that the endpoints are taken such that $a_k = a_{s-k}$ for $k > s/2$.

LEMMA 2.2. *For any adaptive sampling strategy*

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

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

In particular, for standard quickselect ($t = 0$) and quickselect with median-of-$(2t+1)$, we have $v_t(0) = v_t(1) = \frac{2}{3t+4}$, since $r_0 = t+1$ and $s = 2t+1$. For proportion-from-s, $v(0) = v(1) = \frac{s+1}{s^2(s+3)} \sim s^{-2} + O(s^{-3})$. So proportion-from-$s$ has smaller variance when locating elements of either low or high rank than median-of-$(2t+1)$. Furthermore, this result indicates that using large samples can reduce the order of magnitude of the variance; notice that for both types of strategies the coefficient of n^2 in the variance tends to 0 when the size of the samples tends to ∞ and we look for extreme ranks. We will show later that this is indeed true for all ranks.

We now state one of the important results of this paper, where we transform the original problem to one of solving linear differential equations; we arrive at this result after long and careful computations largely similar to those which yield Lemma 2.1.

LEMMA 2.3. *For any adaptive sampling strategy,*

$$\begin{aligned}\frac{d^{s+3}}{d\alpha^{s+3}} g_k(\alpha) &= 2\frac{d^{s+3}}{d\alpha^{s+3}} f_k(\alpha) \\ &+ \frac{(-1)^{s+1-r_k}}{\alpha^{s+1-r_k}} \cdot \frac{s!}{(r_k-1)!} \cdot \frac{d^{r_k+2}}{d\alpha^{r_k+2}} g_k(\alpha) \\ &+ \frac{1}{(1-\alpha)^{r_k}} \cdot \frac{s!}{(s-r_k)!} \cdot \frac{d^{s+3-r_k}}{d\alpha^{s+3-r_k}} g_k(\alpha),\end{aligned}$$

where $g(\alpha)$ *is the second factorial moment characteristic function,* g_k *is its restriction to the* k*th interval,* $f(\alpha)$ *is the expectation characteristic function, and* $\alpha \in I_k$*,* $1 \le k \le \ell$*.*

Thus, except for the independent term $2f^{(s+3)}(\alpha)$ and the higher order derivatives involved, we have the same differential equation as for the expectation characteristic function.

3 The variance of median-of-three

In the case of median-of-three ($s = 3$, $\ell = 1$ and $r_1 = 2$) we are specially lucky, since its expectation characteristic function is $m_1(\alpha) = 2 + 3\alpha(1-\alpha)$ and its sixth derivative vanishes in the differential equation satisfied by $g(\alpha)$ (Lemma 2.3). Hence the corresponding differential equation for $g(\alpha)$ is exactly the same as for the expectation characteristic function $m_1(\alpha)$, namely

$$\frac{d^2\phi}{d\alpha^2} - 6\left(\frac{1}{\alpha^2} + \frac{1}{(1-\alpha)^2}\right)\phi(\alpha) = 0,$$

but here $\phi(\alpha) = g^{(iv)}(\alpha)$, instead of $\phi(\alpha) = m_1'''(\alpha)$.

Hence, the solution has to be integrated four times to recover $g(\alpha)$. The symmetry $g^{(4)}(\alpha) = g^{(4)}(1-\alpha)$ can be used to show that

$$\begin{aligned}\phi(\alpha) &= g^{(4)}(\alpha) \\ &= C_1 \frac{-1 + 2\alpha - 28\alpha^5 + 56\alpha^6 - 40\alpha^7 + 10\alpha^8}{2\alpha^2(1-\alpha)^2}.\end{aligned}$$

Integrating this four times we get four additional arbitrary constants. The technique that we shall use to obtain their value is to plug the general form of $g(\alpha)$ back into the integral equation and compare coefficients. This involves a somewhat long and tedious computation, but it is mostly mechanical and the use of a computer algebra systems is of great help. We finally get

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

and from there

(3.3)

$$\begin{aligned}v_1(\alpha) &= \lim_{n\to\infty, m/n\to\alpha} \mathbb{V}\left[C^{(1)}_{n,m}\right]/n^2 = g(\alpha) - m_1^2(\alpha) \\ &= -\frac{288}{35}\log(1-\alpha) + \frac{576}{35}\alpha\log(1-\alpha) + \frac{2}{7} \\ &\quad - \frac{288}{35}\alpha - \frac{288}{35}\alpha^2(\log\alpha + \log(1-\alpha)) \\ &\quad - \frac{24}{245}\alpha^8 + \frac{96}{245}\alpha^7 - \frac{48}{175}\alpha^6 - \frac{96}{175}\alpha^5 \\ &\quad - \frac{363}{35}\alpha^4 + \frac{774}{35}\alpha^3 - \frac{3657}{1225}\alpha^2,\end{aligned}$$

since $m_1(\alpha) = 2 + 3\alpha(1-\alpha)$. The function $v_1(\alpha)$ is symmetric and has two global maxima at $\alpha \doteq 0.263338\ldots$ and $\alpha \doteq 0.736661\ldots$, where it attains the value $v_1 \doteq 0.391151\ldots$. The global minima are $\alpha = 0$ and $\alpha = 1$, where $v_1 = 2/7 \doteq 0.285714\ldots$, while $v_0(0) = v_0(1) = 1/2$. The function is depicted in Figure 1. A plot of the function $v_0(\alpha)$ corresponding to standard quickselect is shown in Figure 2 for comparison.

The leading coefficient of the variance to locate the median of an array is given by $v_1(1/2) = \frac{144}{35}\log 2 - \frac{97121}{39200} \doteq 0.374229\ldots$. Compare this to $v_0(1/2) = -4\log^2 2 + 4\log 2 + \frac{7}{4} - \frac{\pi^2}{6} \doteq 0.955842\ldots$, where $v_0(\alpha)$ has its unique global maximum.

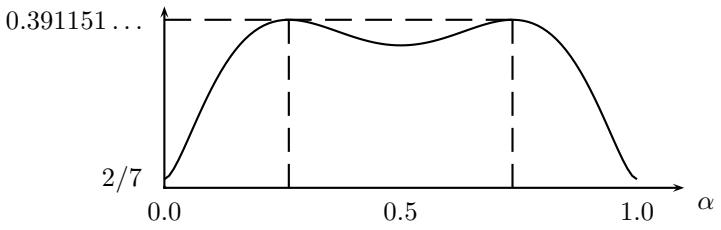

Figure 1: Plot of the coefficient $v_1(\alpha)$ of the variance of quickselect with median-of-three, when looking for the $m = \alpha \cdot n$ smallest element and $n \to \infty$.

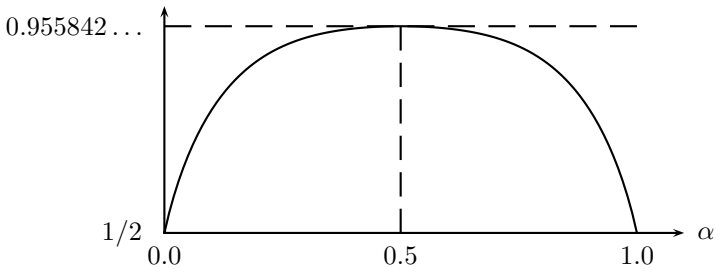

Figure 2: Plot of the coefficient $v_0(\alpha)$ of the variance of standard quickselect, when looking for the $m = \alpha \cdot n$ smallest element and $n \to \infty$ (after [6]).

4 Optimal sampling

Consider a family of *biased* symmetric proportion-from-s sampling strategies, $s \in \mathbb{N}$, with $r(\alpha)/s \to \alpha$ as $s \to \infty$. A sampling strategy is called *biased* [9] if $r(\alpha) > \alpha \cdot s + 1 - \alpha$ for $\alpha < 1/2$. The quantity $\delta = r - s\alpha - 1 + \alpha > 0$ is called the bias. For a biased proportion-from-s strategy the endpoints of the intervals are not evenly distributed in $[0, 1]$ but shifted towards the left for $\alpha < 1/2$ and, symmetrically, towards the right when $\alpha > 1/2$. In [9], it has been shown that optimal average performance is achieved when $s \to \infty$ for such a family. In particular, the limiting expectation characteristic function is $f_\infty(\alpha) = \lim_{s\to\infty} f_s(\alpha) = 1 + \min(\alpha, 1 - \alpha)$.

To tackle a similar analysis for the variance we have to study the behavior of

$$\begin{aligned} T^{(\infty)}(F, G)(\alpha) &= 2F(\alpha) - 1 \\ &\quad + \lim_{s\to\infty} \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \\ &\quad \left\{ \int_\alpha^1 G\left(\frac{\alpha}{x}\right) x^{r(\alpha)+1}(1-x)^{s-r(\alpha)}\, dx \right. \\ &\quad \left. + \int_0^\alpha G\left(\frac{\alpha - x}{1-x}\right) x^{r(\alpha)-1}(1-x)^{s+2-r(\alpha)}\, dx \right\}. \end{aligned}$$

Since $T^{(\infty)}(f_\infty, \cdot)$ is a contraction (see [4, 9]) it suffices to find a fixed point g_∞; if we find it then it is unique, and it is the limit characteristic function for the second factorial moment of our family of sampling strategies. In order to do that, we will need to analyze the asymptotic behavior of

$$I(s) = \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \int_a^b y(x) \cdot x^{r(\alpha)-1} \cdot (1-x)^{s-r(\alpha)}\, dx, \tag{4.4}$$

as $s \to \infty$, as the operator $T^{(\infty)}$ can be expressed as a combination of integrals with the form above, for suitable choices of the integral limits and the function $y(x)$.

We apply Laplace's method to find asymptotic estimations of the value of those integrals (see for instance [2, Ch. 5, p. 211–212]). The maximum of the "kernel" $x^{r-1}(1-x)^{s-r}$ occurs at $x^* = (r-1)/(s-1)$ and so we have two fundamental situations: either the maximum x^* is inside (a, b) and then $I(s) = y(x^*) + \mathcal{O}(s^{-1})$; otherwise, if $x^* \notin [a, b]$ then $I(s) = \mathcal{O}(s^{-1})$ (in general, $I(s) \to 0$ exponentially fast, except if the maximum c were $c = a$ or $c = b$ and $c \sim x^*$). Take $g_\infty(\alpha) = (1 + \min(\alpha, 1-\alpha))^2$. Then

$$\begin{aligned} T^{(\infty)}(f_\infty, g_\infty)(\alpha) &= 2(1 + \min(\alpha, 1-\alpha)) - 1 \\ &\quad + \lim_{s\to\infty} \frac{s!}{(r(\alpha)-1)!(s-r(\alpha))!} \times \Bigg[\\ &\quad \left\{ \int_\alpha^1 x^2 \left(1 + \min\left(\frac{\alpha}{x}, \frac{x-\alpha}{x}\right)\right)^2 \right. \\ &\quad \left. \cdot\, x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}\, dx \right\} \\ &\quad + \left\{ \int_0^\alpha (1-x)^2 \left(1 + \min\left(\frac{\alpha - x}{1-x}, \frac{1-\alpha}{1-x}\right)\right)^2 \right. \\ &\quad \left. \cdot\, x^{r(\alpha)-1}(1-x)^{s-r(\alpha)}\, dx \right\} \Bigg]. \end{aligned}$$

Now, if $\alpha < 1/2$ then, by the definition of a biased strategy, $\alpha < x^*$ and thus the second integral tends to 0, while the first yields the main contribution with $y(x) = (2x - \alpha)^2$. Also, as $x^* = \alpha + \mathcal{O}(s^{-1})$, we have $T^{(\infty)}(g_\infty)(\alpha) \sim 2(1+\alpha) - 1 + \alpha^2 = (1+\alpha)^2$. Because of the symmetry of r, and after careful checking of the special case $\alpha = 1/2$, we finally arrive at

$$T^{(\infty)}(f_\infty, g_\infty) = g_\infty = (1 + \min(\alpha, 1-\alpha))^2.$$

That is, g_∞ is the limiting behavior of the second moment factorial characteristic function of any family of biased proportion-from-s sampling strategies; and $g_\infty = f_\infty^2$. Thus we obtain the following important result.

THEOREM 4.1. *For any family of symmetric* biased *sampling strategies such that* $\lim_{s\to\infty} r(\alpha)/s = \alpha$, *the variance of quickselect using the family of sampling strategies is subquadratic. Indeed,*

$$\lim_{s\to\infty} \lim_{n\to\infty, m/n\to\alpha} \frac{\mathbb{V}[C_{n,m}]}{n^2} = 0.$$

Despite median-of-$(2t+1)$ sampling doesn't yield optimal average behavior, an analogous to Theorem 4.1 holds. In particular, if $t \to \infty$ then the expectation characteristic function $m_t(\alpha) \to 2$ [4]. Using the same techniques that we have just used, we can also easily show that $g_t(\alpha) \to 4$. Hence, the coefficient of n^2 in $\mathbb{V}\left[C^{(t)}_{n,m}\right]$ vanishes as $t \to \infty$. The same type of result was already established for the variance of the cost of selecting an element of random rank by Martínez and Roura [10].

Now we turn our attention to the case of variable-sized samples, i.e., when $s = s(n)$. We assume that $s \to \infty$ as $n \to \infty$ but that it does grow sublinearly ($s = o(n)$). To begin with, the proof given in [9] that the expectation characteristic function of biased proportion-from-s strategies tends to $f_\infty(\alpha) = 1 + \min(\alpha, 1-\alpha)$ as $s \to \infty$ is valid for variable-size samples. The average number of comparisons $\mathbb{E}[C_{n,m}]$ satisfies the recurrence

$$\mathbb{E}[C_{n,m}] = n + \beta\cdot s + o(s) + \sum_{j=m+1}^{n} \pi^{(s,r)}_{n,j} \cdot \mathbb{E}[C_{j-1,m}] + \sum_{j=1}^{m-1} \pi^{(s,r)}_{n,j} \cdot \mathbb{E}[C_{n-j,m-j}]\,;$$

notice the the terms $\beta\cdot s + o(s)$, which account for the average number of comparisons invested in selecting the pivots from the samples. Here, the factor $\beta = \beta(\alpha)$ is the coefficient of the linear cost of the algorithm used to select the pivot. For instance, $\beta(\alpha) = 2 + 3\alpha(1-\alpha)$ if we were using quickselect with median-of-three for selecting pivots from the samples. The full details are not straightforward, but the intuition is rather simple; since $s/n \to 0$ and the asymptotic estimate for the splitting probabilities $\pi^{(s,r)}_{n,j}$ are valid for $s = s(n)$, Theorems 2.1 and 2.2 hold too in this case.

By computing more precise asymptotic estimates of $I(s)$ (see (4.4)), we can establish the behavior of the lower order terms of f_∞ and g_∞. In particular, $f_\infty(\alpha) = 1 + \min(\alpha, 1-\alpha) + \mathcal{O}(s^{-1})$ and $g_\infty(\alpha) = (1+\min(\alpha,1-\alpha))^2 + \mathcal{O}(s^{-1})$. Then $\mathbb{E}[C_{n,m}] = n(1 + \min(\alpha,1-\alpha)) + \beta\cdot s + \mathcal{O}(n/s)$ and $\mathbb{E}\left[C^2_{n,m}\right] = n^2(1+\min(\alpha,1-\alpha))^2 + \Theta\left(n\cdot s + n^2/s\right)$. Hence, we have the following result.

THEOREM 4.2. *Consider a symmetric* biased *sampling strategy with* $s = s(n) \to \infty$ *as* $n \to \infty$, *with* $s = o(n)$ *and such that* $\lim_{s\to\infty} r(\alpha)/s = \alpha$. *Then the variance of quickselect using this sampling strategy is*

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

The variance of a sampling strategy satisfying the hypothesis of the theorem above is minimized when $s = \Theta(\sqrt{n})$. Notice that this is consistent with Theorem 4.1 and that we have then $\sqrt{\mathbb{V}[C_{n,m}]} = \Theta(n^{3/4})$. We prove thus correct several conjectures made in [9] concerning the variance of quickselect with proportion-from-s sampling for variable-sized samples. Also, as $\mathbb{E}[C_{n,m}] = n(1+\min(\alpha,1-\alpha)) + \beta\cdot s + O(n/s)$, it follows that $s = \Theta(\sqrt{n})$ minimizes the average number of comparisons made. Lacking of a better estimate for the term $O(n/s)$ we cannot obtain a more precise asymptotic estimate for the optimal sample size.

References

[1] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publ., New York, 1964.

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

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

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

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

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

[7] 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.

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

[9] 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.

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