Analysis of the Expected Number of Bit Comparisons Required by Quickselect*

James Allen Fill† Takéhiko Nakama‡

Abstract

When algorithms for sorting and searching are applied to keys that are represented as bit strings, we can quantify the performance of the algorithms not only in terms of the number of key comparisons required by the algorithms but also in terms of the number of bit comparisons. Some of the standard sorting and searching algorithms have been analyzed with respect to key comparisons but not with respect to bit comparisons. In this extended abstract, we investigate the expected number of bit comparisons required by `Quickselect` (also known as `Find`). We develop exact and asymptotic formulae for the expected number of bit comparisons required to find the smallest or largest key by `Quickselect` and show that the expectation is asymptotically linear with respect to the number of keys. Similar results are obtained for the average case. For finding keys of arbitrary rank, we derive an exact formula for the expected number of bit comparisons that (using rational arithmetic) requires only finite summation (rather than such operations as numerical integration) and use it to compute the expectation for each target rank.

1 Introduction and Summary

When an algorithm for sorting or searching is analyzed, the algorithm is usually regarded either as comparing keys pairwise irrespective of the keys' internal structure or as operating on representations (such as bit strings) of keys. In the former case, analyses often quantify the performance of the algorithm in terms of the number of key comparisons required to accomplish the task; `Quickselect` (also known as `Find`) is an example of those algorithms that have been studied from this point of view. In the latter case, if keys are represented as bit strings, then analyses quantify the performance of the algorithm in terms of the number of bits compared until it completes its task. Digital search trees, for example, have been examined from this perspective.

In order to fully quantify the performance of a sorting or searching algorithm and enable comparison between key-based and digital algorithms, it is ideal to analyze the algorithm from both points of view. However, to date, only `Quicksort` has been analyzed with both approaches; see Fill and Janson [3]. Before their study, `Quicksort` had been extensively examined with regard to the number of key comparisons performed by the algorithm (e.g., Knuth [11], Régnier [16], Rösler [17], Knessl and Szpankowski [9], Fill and Janson [2], Neininger and Rüschendorf [15]), but it had not been examined with regard to the number of bit comparisons in sorting keys represented as bit strings. In their study, Fill and Janson assumed that keys are independently and uniformly distributed over (0,1) and that the keys are represented as bit strings. [They also conducted the analysis for a general absolutely continuous distribution over (0,1).] They showed that the expected number of bit comparisons required to sort n keys is asymptotically equivalent to $n(\ln n)(\lg n)$ as compared to the lead-order term of the expected number of *key* comparisons, which is asymptotically $2n \ln n$. We use ln and lg to denote natural and binary logarithms, respectively, and use log when the base does not matter (for example, in remainder estimates).

In this extended abstract, we investigate the expected number of bit comparisons required by `Quickselect`. Hoare [7] introduced this search algorithm, which is treated in most textbooks on algorithms and data structures. `Quickselect` selects the m-th smallest key (we call it the rank-m key) from a set of n distinct keys. (The keys are typically assumed to be distinct, but the algorithm still works—with a minor adjustment—even if they are not distinct.) The algorithm finds the target key in a recursive and random fashion. First, it selects a pivot uniformly at random from n keys. Let k denote the rank of the pivot. If $k = m$, then the algorithm returns the pivot. If $k > m$, then the algorithm recursively operates on the set of keys smaller than the pivot and returns the rank-m key. Similarly, if $k < m$, then the algorithm recursively oper-

*Supported by NSF grant DMS–0406104, and by The Johns Hopkins University's Acheson J. Duncan Fund for the Advancement of Research in Statistics.

†Department of Applied Mathematics and Statistics at The Johns Hopkins University.

‡Department of Applied Mathematics and Statistics at The Johns Hopkins University.

ates on the set of keys larger than the pivot and returns the $(k-m)$-th smallest key from the subset. Although previous studies (e.g., Knuth [10], Mahmoud *et al.* [13], Grübel and U. Rösler [6], Lent and Mahmoud [12], Mahmoud and Smythe [14], Devroye [1], Hwang and Tsai [8]) examined `Quickselect` with regard to key comparisons, this study is the first to analyze the bit complexity of the algorithm.

We suppose that the algorithm is applied to n distinct keys that are represented as bit strings and that the algorithm operates on individual bits in order to find a target key. We also assume that the n keys are uniformly and independently distributed in $(0,1)$. For instance, consider applying `Quickselect` to find the smallest key among three keys k_1, k_2, and k_3 whose binary representations are .01001100..., .00110101..., and .00101010..., respectively. If the algorithm selects k_3 as a pivot, then it compares each of k_1 and k_2 to k_3 in order to determine the rank of k_3. When k_1 and k_3 are compared, the algorithm requires 2 bit comparisons to determine that k_3 is smaller than k_1 because the two keys have the same first digit and differ at the second digit. Similarly, when k_2 and k_3 are compared, the algorithm requires 4 bit comparisons to determine that k_3 is smaller than k_2. After these comparisons, key k_3 has been identified as smallest. Hence the search for the smallest key requires a total of 6 bit comparisons (resulting from the two key comparisons).

We let $\mu(m,n)$ denote the expected number of bit comparisons required to find the rank-m key in a file of n keys by `Quickselect`. By symmetry, $\mu(m,n) = \mu(n+1-m,n)$. First, we develop exact and asymptotic formulae for $\mu(1,n) = \mu(n,n)$, the expected number of bit comparisons required to find the smallest key by `Quickselect`, as summarized in the following theorem.

THEOREM 1.1. *The expected number $\mu(1,n)$ of bit comparisons required by* `Quickselect` *to find the smallest key in a file of n keys that are independently and uniformly distributed in $(0,1)$ has the following exact and asymptotic expressions:*

$$\begin{aligned}\mu(1,n) &= 2n(H_n-1) + 2\sum_{j=2}^{n-1} B_j \frac{n-j+1-\binom{n}{j}}{j(j-1)(1-2^{-j})} \\ &= cn - \frac{1}{\ln 2}(\ln n)^2 - \left(\frac{2}{\ln 2}+1\right)\ln n + O(1),\end{aligned}$$

where H_n and B_j denote harmonic and Bernoulli numbers, respectively, and, with $\chi_k := \frac{2\pi i k}{\ln 2}$ and $\gamma :=$ Euler's constant $\doteq 0.57722$, we define

$$\begin{aligned} c &:= \frac{28}{9} + \frac{17-6\gamma}{9\ln 2} \\ &\quad - \frac{4}{\ln 2}\sum_{k\in\mathbb{Z}\setminus\{0\}} \frac{\zeta(1-\chi_k)\Gamma(1-\chi_k)}{\Gamma(4-\chi_k)(1-\chi_k)} \\ &\doteq 5.27938. \end{aligned} \tag{1.1}$$

The constant c can alternatively be expressed as

$$c = 2\sum_{k=0}^{\infty}\left(1 + 2^{-k}\sum_{j=1}^{2^k}\ln\frac{j}{2^k}\right). \tag{1.2}$$

It is easily seen that the expression (1.1) is real, even though it involves the imaginary numbers χ_k. The asymptotic formula shows that the expected number of bit comparisons is asymptotically linear in n with the lead-order coefficient approximately equal to 5.27938. Hence the expected number of bit comparisons is asymptotically different from that of *key* comparisons required to find the smallest key only by a constant factor (the expectation for key comparisons is asymptotically $2n$). Complex-analytic methods are utilized to obtain the asymptotic formula; in a future paper, it will be shown how the linear lead-order asymptotics $\mu(1,n) \sim cn$ [with c given in the form (1.2)] can be obtained without resort to complex analysis. An outline of the proof of Theorem 1.1 is provided in Section 3.

We also derive exact and asymptotic expressions for the expected number of bit comparisons for the average case. We denote this expectation by $\mu(\bar{m},n)$. In the average case, the parameter m in $\mu(m,n)$ is considered a discrete uniform random variable; hence $\mu(\bar{m},n) = \frac{1}{n}\sum_{m=1}^{n}\mu(m,n)$. The derived asymptotic formula shows that $\mu(\bar{m},n)$ is also asymptotically linear in n; see (4.11). More detailed results for $\mu(\bar{m},n)$ are described in Section 4.

Lastly, in Section 5, we derive an exact expression of $\mu(m,n)$ for each fixed m that is suited for computations. Our preliminary exact formula for $\mu(m,n)$ [shown in (2.7)] entails infinite summation and integration. As a result, it is not a desirable form for numerically computing the expected number of bit comparisons. Hence we establish another exact formula that only requires finite summation and use it to compute $\mu(m,n)$ for $m = 1,\ldots,n$, $n = 2,\ldots,25$. The computation leads to the following conjectures: (i) for fixed n, $\mu(m,n)$ [which of course is symmetric about $(n+1)/2$] increases in m for $m \le (n+1)/2$; and (ii) for fixed m, $\mu(m,n)$ increases in n (asymptotically linearly).

Space limitations on this extended abstract force us to omit a substantial portion of the details of our study. We refer the interested reader to our full-length paper [4].

2 Preliminaries

To investigate the bit complexity of `Quickselect`, we follow the general approach developed by Fill and Janson [3]. Let $U_1, \ldots, U_n$ denote the n keys uniformly and independently distributed on (0, 1), and let $U_{(i)}$ denote the rank-i key. Then, for $1 \le i < j \le n$ (assume $n \ge 2$),

$$P\{U_{(i)} \text{ and } U_{(j)} \text{ are compared}\} = \begin{cases} \dfrac{2}{j-m+1} & \text{if } m \le i \\ \dfrac{2}{j-i+1} & \text{if } i < m < j \\ \dfrac{2}{m-i+1} & \text{if } j \le m. \end{cases} \tag{2.1}$$

To determine the first probability in (2.1), note that $U_{(m)}, \ldots, U_{(j)}$ remain in the same subset until the first time that one of them is chosen as a pivot. Therefore, $U_{(i)}$ and $U_{(j)}$ are compared if and only if the first pivot chosen from $U_{(m)}, \ldots, U_{(j)}$ is either $U_{(i)}$ or $U_{(j)}$. Analogous arguments establish the other two cases.

For $0 < s < t < 1$, it is well known that the joint density function of $U_{(i)}$ and $U_{(j)}$ is given by

$$f_{U_{(i)},U_{(j)}}(s,t) := \binom{n}{i-1,1,j-i-1,1,n-j} \times s^{i-1}(t-s)^{j-i-1}(1-t)^{n-j}. \tag{2.2}$$

Clearly, the event that $U_{(i)}$ and $U_{(j)}$ are compared is independent of the random variables $U_{(i)}$ and $U_{(j)}$. Hence, defining

$$P_1(s,t,m,n) := \sum_{m \le i < j \le n} \frac{2}{j-m+1} f_{U_{(i)},U_{(j)}}(s,t), \tag{2.3}$$

$$P_2(s,t,m,n) := \sum_{1 \le i < m < j \le n} \frac{2}{j-i+1} f_{U_{(i)},U_{(j)}}(s,t), \tag{2.4}$$

$$P_3(s,t,m,n) := \sum_{1 \le i < j \le m} \frac{2}{m-i+1} f_{U_{(i)},U_{(j)}}(s,t), \tag{2.5}$$

$$P(s,t,m,n) := P_1(s,t,m,n) + P_2(s,t,m,n) + P_3(s,t,m,n) \tag{2.6}$$

[the sums in (2.3)–(2.5) are double sums over i and j], and letting $\beta(s,t)$ denote the index of the first bit at which the keys s and t differ, we can write the expectation $\mu(m,n)$ of the number of bit comparisons required to find the rank-m key in a file of n keys as

$$\begin{aligned} \mu(m,n) &= \int_0^1 \int_s^1 \beta(s,t) P(s,t,m,n)\, dt\, ds \\ &= \sum_{k=0}^{\infty} \sum_{l=1}^{2^k} \int_{(l-1)2^{-k}}^{(l-\frac{1}{2})2^{-k}} \int_{(l-\frac{1}{2})2^{-k}}^{l2^{-k}} (k+1) \\ &\qquad \times P(s,t,m,n)\, dt\, ds; \end{aligned} \tag{2.7}$$

in this expression, note that k represents the last bit at which s and t agree.

3 Analysis of $\mu(1,n)$

In Section 3.1, we outline a derivation of the exact expression for $\mu(1,n)$ shown in Theorem 1.1; see the full paper [4] for the numerous suppressed details of the various computations. In Section 3.2, we prove the asymptotic result asserted in Theorem 1.1.

3.1 Exact Computation of $\mu(1,n)$ Since the contribution of $P_2(s,t,m,n)$ or $P_3(s,t,m,n)$ to $P(s,t,m,n)$ is zero for $m = 1$, we have $P(s,t,1,n) = P_1(s,t,1,n)$ [see (2.4) through (2.6)]. Let $x := s$, $y := t - s$, $z := 1 - t$. Then

$$\begin{aligned} &P_1(s,t,1,n) \\ &\quad = z^n \sum_{1 \le i < j \le n} \frac{2}{j} \binom{n}{i-1,1,j-i-1,1,n-j} \\ &\qquad \times x^{i-1} y^{j-i-1} z^{-j} \\ &\quad = 2 \sum_{j=2}^{n} (-1)^j \binom{n}{j} t^{j-2}. \end{aligned} \tag{3.1}$$

From (2.7) and (3.1),

$$\mu(1,n) = \sum_{j=2}^{n} \frac{(-1)^j \binom{n}{j}}{j-1} \sum_{k=0}^{\infty} (k+1) 2^{-kj} \sum_{l=1}^{2^k} [l^{j-1} - (l-\tfrac{1}{2})^{j-1}]. \tag{3.2}$$

To further transform (3.2), define

$$a_{j,r} = \begin{cases} \dfrac{B_r}{r} \binom{j-1}{r-1} & \text{if } r \ge 2 \\ \frac{1}{2} & \text{if } r = 1 \\ \frac{1}{j} & \text{if } r = 0, \end{cases} \tag{3.3}$$

where B_r denotes the r-th Bernoulli number. Let $S_{n,j} := \sum_{l=1}^{n} l^{j-1}$. Then $S_{n,j} = \sum_{r=0}^{j-1} a_{j,r} n^{j-r}$ (see Knuth [11]), and

$$\begin{aligned}\mu(1,n) &= 2\sum_{j=2}^{n} \frac{(-1)^j \binom{n}{j}}{j-1} \sum_{k=0}^{\infty} (k+1)2^{-kj} \\ &\quad \times \sum_{r=1}^{j-1} a_{j,r} 2^{k(j-r)}(1-2^{-r}) \\ &= 2n(H_n - 1) + 2t_n, \end{aligned} \tag{3.4}$$

where H_n denotes the n-th harmonic number and

$$t_n := \sum_{j=2}^{n-1} \frac{B_j}{j(1-2^{-j})} \left[\frac{n - \binom{n}{j}}{j-1} - 1 \right]. \tag{3.5}$$

3.2 Asymptotic Analysis of $\mu(1,n)$ In order to obtain an asymptotic expression for $\mu(1,n)$, we analyze t_n in (3.4)–(3.5). The following lemma provides an exact expression for t_n that easily leads to an asymptotic expression for $\mu(1,n)$:

LEMMA 3.1. *Let γ denote Euler's constant ($\doteq 0.57722$), and define $\chi_k := \frac{2\pi i k}{\ln 2}$. Then*

$$\begin{aligned} t_n &= -(nH_n - n - 1) + a(n-2) \\ &\quad - \frac{1}{2\ln 2}\left[H_n^2 + H_n^{(2)} - \frac{7}{2}\right] \\ &\quad + \left(\frac{\gamma - 1}{\ln 2} - \frac{1}{2}\right)\left(H_n - \frac{3}{2}\right) \\ &\quad + b - \Sigma_n, \end{aligned}$$

where

$$\begin{aligned} a &:= \frac{14}{9} + \frac{17 - 6\gamma}{18\ln 2} \\ &\quad - \frac{2}{\ln 2} \sum_{k \in \mathbb{Z}\setminus\{0\}} \frac{\zeta(1-\chi_k)\Gamma(1-\chi_k)}{\Gamma(4-\chi_k)(1-\chi_k)}, \\ b &:= \sum_{k \in \mathbb{Z}\setminus\{0\}} \frac{2\zeta(1-\chi_k)\Gamma(-\chi_k)}{(\ln 2)(1-\chi_k)\Gamma(3-\chi_k)}, \\ \Sigma_n &:= \sum_{k \in \mathbb{Z}\setminus\{0\}} \frac{\zeta(1-\chi_k)\Gamma(-\chi_k)\Gamma(n+1)}{(\ln 2)(1-\chi_k)\Gamma(n+1-\chi_k)}, \end{aligned}$$

and $H_n^{(2)}$ denotes the n-th Harmonic number of order 2, i.e., $H_n^{(2)} := \sum_{i=1}^{n} \frac{1}{i^2}$.

The proof of the lemma involves complex-analytic techniques and is rather lengthy, so it is omitted in this extended abstract; see our full-length paper [4]. From (3.4), the exact expression for t_n also provides an alternative exact expression for $\mu(1,n)$.

Using Lemma 3.1, we complete the proof of Theorem 1.1. We know

$$H_n = \ln n + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + O(n^{-4}), \tag{3.6}$$

$$H_n^{(2)} = \frac{\pi^2}{6} - \frac{1}{n} + \frac{1}{2n^2} + O(n^{-3}). \tag{3.7}$$

Combining (3.6)–(3.7) with (3.4) and Lemma 3.1, we obtain an asymptotic expression for $\mu(1,n)$:

$$\mu(1,n) = 2an - \frac{1}{\ln 2}(\ln n)^2 - \left(\frac{2}{\ln 2} + 1\right)\ln n + O(1). \tag{3.8}$$

The term $O(1)$ in (3.8) has fluctuations of small magnitude due to Σ_n, which is periodic in $\log n$ with amplitude smaller than 0.00110. Thus, as asserted in Theorem 1.1, the asymptotic slope in (3.8) is

$$\begin{aligned} c &= 2a \\ &= \frac{28}{9} + \frac{17 - 6\gamma}{9\ln 2} - \frac{4}{\ln 2} \sum_{k \in \mathbb{Z}\setminus\{0\}} \frac{\zeta(1-\chi_k)\Gamma(1-\chi_k)}{\Gamma(4-\chi_k)(1-\chi_k)}. \end{aligned} \tag{3.9}$$

The alternative expression (1.2) for c is established in a forthcoming revision to our full-length paper [4]; this was also done independently by Grabner and Prodinger [5]. As described in their paper, suitable use of Stirling's formula with bounds allows one to compute c very rapidly to many decimal places.

4 Analysis of the Average Case: $\mu(\bar{m}, n)$

4.1 Exact Computation of $\mu(\bar{m}, n)$ Here we consider the parameter m in $\mu(m,n)$ as a discrete random variable with uniform probability mass function $P\{m = i\} = 1/n$, $i = 1, 2, \ldots, n$, and average over m while the parameter n is fixed. Thus, using the notation defined in Section 2,

$$\mu(\bar{m}, n) = \mu_1(\bar{m}, n) + \mu_2(\bar{m}, n) + \mu_3(\bar{m}, n),$$

where, for $l = 1, 2, 3$,

$$\mu_l(\bar{m}, n) = \int_0^1 \int_s^1 \beta(s,t) \, \frac{1}{n} \sum_{m=1}^{n} P_l(s,t,m,n) \, dt \, ds. \tag{4.1}$$

Here $\mu_1(\bar{m}, n) = \mu_3(\bar{m}, n)$ by an easy symmetric argument we omit, and so

$$\mu(\bar{m}, n) = 2\mu_1(\bar{m}, n) + \mu_2(\bar{m}, n); \tag{4.2}$$

we will compute $\mu_1(\bar{m}, n)$ and $\mu_2(\bar{m}, n)$ exactly in Section 4.1.1.

4.1.1 Exact Computation of $\mu(\bar{m}, n)$ We use the following lemma in order to compute $\mu_1(\bar{m}, n)$ exactly:

LEMMA 4.1.

$$\begin{aligned}
&\int_0^1 \int_s^1 \beta(s,t)\, \frac{1}{n} \sum_{m=2}^{n} P_1(s,t,m,n)\, dt\, ds \\
&= 2\sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{j(j-1)} + \frac{2}{9} \sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{j-1} \\
&\quad -2\sum_{j=3}^{n-1} B_j \frac{n-j+1-\binom{n-1}{j-1}}{j(j-1)(j-2)(1-2^{-j})} \\
&\quad -2\sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{(j+1)j(j-1)(1-2^{-j})}.
\end{aligned}$$

Space limitations on this extended abstract do not allow us to prove this lemma here; we give the proof in our full-length paper [4]. Since

$$\begin{aligned}
\mu_1(\bar{m}, n) &= \frac{1}{n}\mu(1,n) \\
&\quad + \int_0^1 \int_s^1 \beta(s,t)\, \frac{1}{n} \sum_{m=2}^{n} P_1(s,t,m,n)\, dt\, ds,
\end{aligned}$$

it follows from (3.4) and Lemma 4.1 that

$$\begin{aligned}
\mu_1(\bar{m}, n) &= n - 1 - 4\sum_{j=3}^{n} \frac{(-1)^j \binom{n-1}{j-1}}{j(j-1)(j-2)} \\
&\quad + \frac{2}{n} \sum_{j=2}^{n-1} B_j \frac{n-j+1-\binom{n}{j}}{j(j-1)(1-2^{-j})} \\
&\quad + \frac{2}{9} \sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{j-1} \\
&\quad -2\sum_{j=3}^{n-1} B_j \frac{n-j+1-\binom{n-1}{j-1}}{j(j-1)(j-2)(1-2^{-j})} \\
&\quad -2\sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{(j+1)j(j-1)(1-2^{-j})}. \qquad (4.3)
\end{aligned}$$

Similarly, after laborious calculations, one can show that

$$\mu_2(\bar{m}, n) = -\frac{4}{n} \sum_{j=2}^{n} \frac{(-1)^j \binom{n}{j}}{j(j-1)[1-2^{-(j-1)}]} + 2(n-1). \qquad (4.4)$$

From (4.2)–(4.4), we obtain

$$\begin{aligned}
&\mu(\bar{m}, n) \\
&= 2(n-1) - 8\sum_{j=3}^{n} \frac{(-1)^j \binom{n-1}{j-1}}{j(j-1)(j-2)} \\
&\quad + \frac{4}{n} \sum_{j=2}^{n-1} B_j \frac{n-j+1-\binom{n}{j}}{j(j-1)(1-2^{-j})} \\
&\quad + \frac{4}{9} \sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{j-1} \\
&\quad -4\sum_{j=3}^{n-1} B_j \frac{n-j+1-\binom{n-1}{j-1}}{j(j-1)(j-2)(1-2^{-j})} \\
&\quad -4\sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{(j+1)j(j-1)(1-2^{-j})} \\
&\quad -\frac{4}{n} \sum_{j=2}^{n} \frac{(-1)^j \binom{n}{j}}{j(j-1)[1-2^{-(j-1)}]} + 2(n-1). \qquad (4.5)
\end{aligned}$$

We rewrite or combine some of the terms in (4.5) for the asymptotic analysis of $\mu(\bar{m}, n)$ described in the next section. We define

$$\begin{aligned}
F_1(n) &:= \sum_{j=3}^{n} \frac{(-1)^j \binom{n}{j}}{(j-1)(j-2)}, \\
F_2(n) &:= \sum_{j=2}^{n-1} \frac{B_j}{j(1-2^{-j})} \left[\frac{n-\binom{n}{j}}{j-1} - 1 \right], \\
F_3(n) &:= \sum_{j=2}^{n-1} \frac{(-1)^j \binom{n-1}{j}}{j-1} \\
F_4(n) &:= \sum_{j=3}^{n-1} \frac{B_j}{j(j-1)(1-2^{-j})} \left[\frac{n-1-\binom{n-1}{j-1}}{j-2} - 1 \right], \\
F_5(n) &:= \sum_{j=3}^{n} \frac{(-1)^j \binom{n}{j}}{j(j-1)(j-2)[1-2^{-(j-1)}]}.
\end{aligned}$$

Then

$$\begin{aligned}
\mu(\bar{m}, n) &= 2(n-1) - \tfrac{8}{n}F_1(n) + \tfrac{4}{n}F_2(n) + \tfrac{4}{9}F_3(n) \\
&\quad -4F_4(n) + \tfrac{8}{n}F_5(n). \qquad (4.6)
\end{aligned}$$

4.2 Asymptotic Analysis of $\mu(\bar{m}, n)$ We derive an asymptotic expression for $\mu(\bar{m}, n)$ shown in (4.6).

Routine arguments show that

$$F_1(n) = -\frac{1}{2}n^2\ln n + \left(\frac{5}{4}-\frac{\gamma}{2}\right)n^2 - n\ln n + \frac{n^2}{2(n-1)} - (\gamma+1)n + O(1), \tag{4.7}$$

$$F_3(n) = n\ln n + (\gamma-1)n - \ln n + O(1), \tag{4.8}$$

$$F_4(n) = \frac{1}{9}n\ln n + \left(\tilde{a}+\frac{1}{9}\gamma-\frac{1}{9}\right)n + \frac{8}{9}\ln n + O(1), \tag{4.9}$$

$$F_5(n) = -\frac{1}{2}n^2\ln n + \frac{3+\ln 2-\gamma}{2}n^2 - \frac{1}{2\ln 2}n(\ln n)^2 + \left(\frac{1}{\ln 2}-\frac{1}{2}\right)n\ln n + O(n), \tag{4.10}$$

where

$$\tilde{a} := \frac{7}{36\ln 2} - \frac{41}{72} - \frac{\gamma}{12\ln 2} - \sum_{k\in Z\setminus\{0\}} \frac{\zeta(1-\chi_k)\Gamma(1-\chi_k)}{(\ln 2)(2-\chi_k)\Gamma(4-\chi_k)}.$$

Since $F_2(n)$ is equal to t_n, which is defined at (3.5) and analyzed in Section 3.2, we already have an asymptotic expression for $F_2(n)$. Therefore, from (4.6)–(4.10), we obtain the following asymptotic formula for $\mu(\bar{m}, n)$:

$$\mu(\bar{m},n) = 4(1+\ln 2-\tilde{a})n - \frac{4}{\ln 2}(\ln n)^2 + 4\left(\frac{2}{\ln 2}-1\right)\ln n + O(1). \tag{4.11}$$

The asymptotic slope $4(1+\ln 2-\tilde{a})$ is approximately 8.20731. We have not (yet) sought an alternative form for $\tilde{a}$ like that for c in (1.2).

5 Derivation of a Closed Formula for $\mu(m,n)$

The exact expression for $\mu(m,n)$ obtained in Section 2 [see (2.7)] involves infinite summation and integration. Hence it is not a preferable form for numerically computing the expectation. In this section, we establish another exact expression for $\mu(m,n)$ that only involves finite summation. We also use the formula to compute $\mu(m,n)$ for $m = 1,\ldots,n$, $n = 2,\ldots,20$.

As described in Section 2, it follows from equations (2.6)–(2.7) that

$$\mu(m,n) = \mu_1(m,n) + \mu_2(m,n) + \mu_3(m,n),$$

where, for $q = 1, 2, 3$,

$$\mu_q(m,n) := \sum_{k=0}^{\infty}\sum_{l=1}^{2^k}\int_{s=(l-1)2^{-k}}^{(l-\frac{1}{2})2^{-k}}\int_{t=(l-\frac{1}{2})2^{-k}}^{l2^{-k}}(k+1)\times P_q(s,t,m,n)\,dt\,ds. \tag{5.1}$$

The same technique can be applied to eliminate the infinite summation and integration from each $\mu_q(m,n)$. We describe the technique for obtaining a closed expression of $\mu_1(m,n)$.

First, we transform $P_1(s,t,m,n)$ shown in (2.3) so that we can eliminate the integration in $\mu_1(m,n)$. Define

$$C_1(i,j) := I\{1\le m\le i<j\le n\} \times \frac{2}{j-m+1}\binom{n}{i-1,1,j-i-1,1,n-j}, \tag{5.2}$$

where $I\{1 \le m \le i < j \le n\}$ is an indicator function that equals 1 if the event in braces holds and 0 otherwise. Then

$$P_1(s,t,m,n) = \sum_{f=m-1}^{n-2}\sum_{h=0}^{n-f-2} s^f t^h C_2(f,h), \tag{5.3}$$

where

$$C_2(f,h) := \sum_{i=m}^{f+1}\sum_{j=f+2}^{f+h+2} C_1(i,j)\binom{j-i-1}{f-i+1} \times \binom{n-j}{h-j+f+2}(-1)^{h-i-j+1}.$$

Thus, from (5.1) and (5.3), we can eliminate the integration in $\mu_1(m,n)$ and express it using polynomials in l:

$$\mu_1(m,n) = \sum_{f=m-1}^{n-2}\sum_{h=0}^{n-f-2} C_3(f,h)\sum_{k=0}^{\infty}(k+1) \times \sum_{l=1}^{2^k} 2^{-k(f+h+2)}[l^{h+1} - (l-\tfrac{1}{2})^{h+1}] \times [(l-\tfrac{1}{2})^{f+1} - (l-1)^{f+1}], \tag{5.4}$$

where

$$C_3(f,h) := \frac{1}{(n+1)(f+1)}C_2(f,h).$$

One can show that

$$\left[l^{h+1} - \left(l - \frac{1}{2}\right)^{h+1}\right]\left[\left(l - \frac{1}{2}\right)^{f+1} - (l-1)^{f+1}\right]$$

$$= \sum_{j=1}^{f+h+1} C_4(f,h,j) l^{j-1}, \tag{5.5}$$

where

$$C_4(f,h,j) := (-1)^{f+h-j+1}\left(\frac{1}{2}\right)^{h-j+2} \times \sum_{j'=0 \bigvee (j-1-h)}^{(j-1)\bigwedge f} \binom{f+1}{j'}\binom{h+1}{j-1-j'} \times \left[1 - \left(\frac{1}{2}\right)^{f+1-j'}\right]\left(\frac{1}{2}\right)^{j'}.$$

From (5.4)–(5.5), we obtain

$$\mu_1(m,n) = \sum_{f=m-1}^{n-2}\sum_{h=0}^{n-f-2}\sum_{j=1}^{f+h+1} C_5(f,h,j) \times \sum_{k=0}^{\infty}(k+1)2^{-k(f+h+2)}\sum_{l=1}^{2^k} l^{j-1},$$

where

$$C_5(f,h,j) := C_3(f,h) \cdot C_4(f,h,j).$$

Here, as described in Section 3.1,

$$\sum_{l=1}^{2^k} l^{j-1} = \sum_{r=0}^{j-1} a_{j,r} 2^{k(j-r)},$$

where $a_{j,r}$ is defined by (3.3). Now define

$$C_6(f,h,j,r) := a_{j,r}\ C_5(f,h,j).$$

Then

$$\mu_1(m,n) = \sum_{a=1}^{n-1} C_7(a)(1-2^{-a})^{-2}, \tag{5.6}$$

where

$$C_7(a) := \sum_{f=m-1}^{n-2}\sum_{h=\alpha}^{n-f-2}\sum_{j=\beta}^{f+h+1} C_6(f,h,j,a+j-(f+h+2)),$$

in which $\alpha := 0 \bigvee (a-f-1)$ and $\beta := 1 \bigvee (f+h+2-a)$.

The procedure described above can be applied to derive analogous exact formulae for $\mu_2(m,n)$ and $\mu_3(m,n)$. In order to derive the analogous exact formula for $\mu_2(m,n)$, one need only start the derivation by changing the indicator function in $C_1(i,j)$ [see (5.2)] to $I\{1 \le i < m < j \le n\}$ and follow each step of the procedure; similarly, for $\mu_3(m,n)$, one need only start the derivation by changing the indicator function to $I\{1 \le i < j \le m \le n\}$.

Using the closed exact formulae of $\mu_1(m,n)$, $\mu_2(m,n)$, and $\mu_3(m,n)$, we computed $\mu(m,n)$ for $n = 2,3,\ldots,20$ and $m = 1,2,\ldots,n$. Figure 1 shows the results, which suggest the following: (i) for fixed n, $\mu(m,n)$ [which of course is symmetric about $(n+1)/2$] increases in m for $m \le (n+1)/2$; and (ii) for fixed m, $\mu(m,n)$ increases in n (asymptotically linearly).

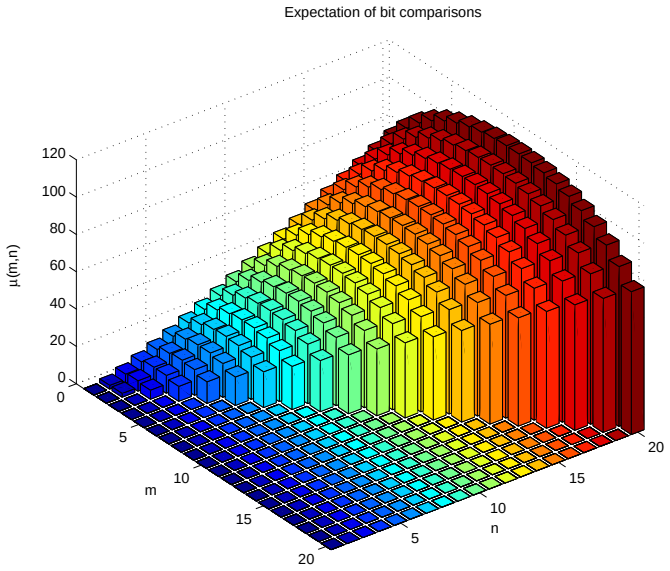

Figure 1: Expected number of bit comparisons for `Quickselect`. The closed formulae for $\mu_1(m,n)$, $\mu_2(m,n)$, and $\mu_3(m,n)$ were used to compute $\mu(m,n)$ for $n = 1,2,\ldots,20$ (n represents the number of keys) and $m = 1,2,\ldots,n$ (m represents the rank of the target key).

6 Discussion

Our investigation of the bit complexity of `Quickselect` revealed that the expected number of bit comparisons required by `Quickselect` to find the smallest or largest key from a set of n keys is asymptotically linear in n with the asymptotic slope approximately equal to

5.27938. Hence asymptotically it differs from the expected number of *key* comparisons to achieve the same task only by a constant factor. (The expectation for key comparisons is asymptotically $2n$; see Knuth [10] and Mahmoud *et al.* [13]). This result is rather contrastive to the `Quicksort` case in which the expected number of bit comparisons is asymptotically $n(\ln n)(\lg n)$ whereas the expected number of key comparisons is asymptotically $2n \ln n$ (see Fill and Janson [3]). Our analysis also showed that the expected number of bit comparisons for the average case remains asymptotically linear in n with the lead-order coefficient approximately equal to 8.20731. Again, the expected number is asymptotically different from that of key comparisons for the average case only by a constant factor. (The expected number of key comparisons for the average case is asymptotically $3n$; see Mahmoud *et al.* [13]).

Although we have yet to establish a formula analogous to (3.4) and (4.6) for the expected number of bit comparisons to find the m-th key for fixed m, we established an exact expression that only requires finite summation and used it to obtain the results shown in Figure 1. However, the formula remains computationally complex. Written as a single expression, $\mu(m, n)$ is a seven-fold sum of rather elementary terms with each sum having order n terms (in the worst case); in this sense, the running time of the algorithm for computing $\mu(m, n)$ is of order n^7. The expression for $\mu(m, n)$ does not allow us to derive an asymptotic formula for it or to prove the two intuitively obvious observations described at the end of Section 5. The situation is substantially better for the expected number of *key* comparisons to find the m-th key from a set of n keys; Knuth [10] showed that the expectation can be written as $2[n+3+(n+1)H_n-(m+2)H_m-(n+3-m)H_{n+1-m}]$.

In this extended abstract, we considered independent and uniformly distributed keys in (0,1). In this case, each bit in a bit-string key is 1 with probability 0.5. In ongoing research, we generalize the model and suppose that each bit results from an independent Bernoulli trial with success probability p. The more general results of that research will further elucidate the bit complexity of `Quickselect` and other algorithms.

Acknowledgment. We thank Philippe Flajolet, Svante Janson, and Helmut Prodinger for helpful discussions.

References

[1] L. Devroye. On the probablistic worst-case time of "Find". *Algorithmica*, 31:291–303, 2001.

[2] J. A. Fill and S. Janson. Quicksort asymptotics. *Journal of Algorithms*, 44:4–28, 2002.

[3] J. A. Fill and S. Janson. The number of bit comparisons used by Quicksort: An average-case analysis. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 293–300, 2004.

[4] J. A. Fill and T. Nakama. Analysis of the expected number of bit comparisons required by Quickselect. *http://front.math.ucdavis.edu/0706.2437*, 2007.

[5] P. J. Grabner and H. Prodinger. On a constant arising in the analysis of bit comparisons in Quickselect. *Preprint*, 2007.

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

[7] C. R. Hoare. Find (algorithm 65). *Communications of the ACM*, 4:321–322, 1961.

[8] H. Hwang and T. Tsai. Quickselect and the Dickman function. *Combinatorics, Probability and Computing*, 11:353–371, 2002.

[9] C. Knessl and W. Szpankowski. Quicksort algorithm again revisited. *Discrete Mathematics and Theoretical Computer Science*, 3:43–64, 1999.

[10] D. E. Knuth. Mathematical analysis of algorithms. In *Information Processing 71 (Proceedings of IFIP Congress, Ljubljana, 1971)*, pages 19–27. North-Holland, Amsterdam, 1972.

[11] D. E. Knuth. *The Art of Computer Programming. Volume 3: Sorting and Searching*. Addison-Wesley, Reading, Massachusetts, 1998.

[12] J. Lent and H. M. Mahmoud. Average-case analysis of multiple Quickselect: An algorithm for finding order statistics. *Statistics and Probability Letters*, 28:299–310, 1996.

[13] H. M. Mahmoud, R. Modarres, and R. T. Smythe. Analysis of Quickselect: An algorithm for order statistics. *RAIRO Informatique Théorique et Applications*, 29:255–276, 1995.

[14] H. M. Mahmoud and R. T. Smythe. Probabilistic analysis of multiple Quickselect. *Algorithmica*, 22:569–584, 1998.

[15] R. Neininger and L. Rüschendorf. Rates of convergence for Quickselect. *Journal of Algorithm*, 44:51–62, 2002.

[16] M. Régnier. A limiting distribution of Quicksort. *RAIRO Informatique Théorique et Applications*, 23:335–343, 1989.

[17] U. Rösler. A limit theorem for Quicksort. *RAIRO Informatique Théorique et Applications*, 25:85–100, 1991.