Electronic Journal of Probability

Electron. J. Probab. **23** (2018), no. 100, 1–13.
ISSN: 1083-6489 https://doi.org/10.1214/18-EJP227

Non-asymptotic distributional bounds for the Dickman approximation of the running time of the Quickselect algorithm*

Larry Goldstein†

Abstract

Given a non-negative random variable W and $\theta > 0$, let the generalized Dickman transformation map the distribution of W to that of

$$W^* =_d U^{1/\theta}(W+1),$$

where $U \sim \mathcal{U}[0,1]$, a uniformly distributed variable on the unit interval, independent of W, and where $=_d$ denotes equality in distribution. It is well known that W^* and W are equal in distribution if and only if W has the generalized Dickman distribution $\mathcal{D}_\theta$. We demonstrate that the Wasserstein distance d_1 between W, a non-negative random variable with finite mean, and D_θ having distribution $\mathcal{D}_\theta$ obeys the inequality

$$d_1(W, D_\theta) \le (1+\theta) d_1(W, W^*).$$

The specialization of this bound to the case $\theta = 1$ and coupling constructions yield

$$d_1(W_{n,1}, D_1) \le \frac{8\log(n/2)+10}{n} \quad \text{for all } n \ge 1, \text{ where for } m \ge 1 \quad W_{n,m} = \frac{1}{n} C_{n,m} - 1,$$

and $C_{n,m}$ is the number of comparisons made by the Quickselect algorithm to find the m^{th} smallest element of a list of n distinct numbers. A similar bound holds for $W_{n,m}$ for $m \ge 2$, and together recover and quantify the results of [12] that show distributional convergence of $W_{n,m}$ to the standard Dickman distribution $\mathcal{D}_1$ in the asymptotic regime $m = o(n)$. By comparison to an exact expression for the expected running time $E[C_{n,m}]$, lower bounds are provided that show the rate is not improvable for $m \neq 2$.

Keywords: sorting; complexity; distributional approximation.
AMS MSC 2010: 60F05;68Q25.
Submitted to EJP on January 30, 2018, final version accepted on September 21, 2018.
Supersedes arXiv:1703.00505.

*This work was partially supported by NSA grant H98230-15-1-0250.
†Department of Mathematics, University of Southern California. E-mail: larry@math.usc.edu

1 Introduction

For a given non-negative random variable W and $\theta > 0$, let the generalized Dickman transformation map the distribution of W to that of

$$W^* =_d U^{1/\theta}(W+1), \tag{1.1}$$

where U has the uniform distribution $\mathcal{U}[0,1]$ on the unit interval, and is independent of W and where $=_d$ denotes equality in distribution. It is well known [6], [15] that the generalized Dickman distribution $\mathcal{D}_\theta$ is the unique fixed point of the transformation (1.1), that is,

$$W \sim \mathcal{D}_\theta \quad \text{if and only if} \quad W =_d W^*. \tag{1.2}$$

When (1.1) holds we will say that W^* has the $\mathcal{D}_\theta$-bias distribution of W. In what follows, D_θ will denote a random variable with distribution $\mathcal{D}_\theta$. The case $\theta = 1$ corresponds to the (standard) Dickman distribution, for which we may drop the subscript θ.

The Dickman function ρ first made its appearance in number theory [7] when counting the number of integers below a fixed threshold whose prime factors satisfy some given upper bound. Standardizing ρ yields the density of the standard Dickman distribution, the cannonical member of the family $\mathcal{D}_\theta, \theta > 0$ of generalized Dickman distributions, which also arise in the study of component counts of logarithmic combinatorial structures such as permutations and partitions [1], and more generally for the quasi-logarithmic class considered in [3]. See also the recent work [17], [2] and [4] in this area, that detail some connections to probabilistic number theory.

Members from the generalized Dickman family have subsequently been noted to arise in a variety of other contexts, in particular for the sum of edge lengths of vertices connected to the origin in minimal directed spanning trees in [15], and for weighted sums of independent random variables in [16], [2] and [4]. Simulation of the Dickman distribution has been considered in [6].

Here we study the error incurred when using the standard Dickman distribution to approximate that of the (properly normalized) number of comparisons made by the Quickselect sorting algorithm of Hoare [11] for locating the m^{th} smallest element of a list of n distinct numbers. One may visualize how Quickselect works in terms of a tree structure. First, a 'pivot' is chosen uniformly from the given list. The list is then divided into those numbers on the list that are strictly smaller, making up the left subtree, and those that are strictly larger, making up the right. If the left subtree is of size $m-1$ then the pivot is the desired m^{th} smallest element, and the procedure terminates. Otherwise, the process continues recursively on the left sub-tree if it is of size m or larger, and else on the right sub-tree.

Letting

$$W_{n,m} = \frac{1}{n} C_{n,m} - 1, \tag{1.3}$$

where $C_{n,m}$ is the number of comparisons made by Quickselect, the work of [12] showed that $W_{n,m}$ converges in distribution to the Dickman D when $m = o(n)$. We note that in the case $m = 1$ Quickselect simplifies in that at each step of the recursion the procedure either stops or continues on the left subtree. As this case is simpler than for $m \geq 2$ we deal with it separately.

The following two theorems quantify and recover the results of [12] by providing non-asypmptotic bounds in the Wasserstein distance d_1 between $W_{n,m}$ and D that converge to zero in the $m = o(n)$ asymptotic regime. As the m^{th} smallest number of a list of n distinct numbers only exists when $n \geq m$, we need only consider this range of parameters in what follows.

Theorem 1.1. *Let $C_{n,1}$ be the number of comparisons made by Quickselect to find the smallest of a list of n distinct numbers, and let $W_{n,1}$ be given by* (1.3)*. Then for all $n \ge 1$*

$$d_1(W_{n,1}, D) \le \frac{8\log(n/2) + 10}{n}.$$

Theorem 1.2. *Let $m \ge 2$ and $C_{n,m}$ the number of comparisons made by Quickselect to find the m^{th} smallest element of a list of n distinct numbers, and let $W_{n,m}$ be given by* (1.3)*. Then for all $n \ge m$*

$$d_1(W_{n,m}, D) \le \frac{(46m+8)\log(n/m) + 54m + 8}{n}.$$

That the bounds in Theorems 1.1 and 1.2 are tight in the $\log n/n$ order for $m \neq 2$ is a consequence of the following result; in the following, we let $h_n = \sum_{1\le k\le n} 1/k$ for $n \ge 1$.

Theorem 1.3. *For all $m \ge 1$,*

$$d_1(W_{n,m}, D) \ge \frac{2(|m-2|\log n - |(m+2)h_m - 3|)}{n} \quad \textit{for all } n \ge m.$$

We note that in the case $m = 1$ the lower bound simplifies to $2\log n/n$. That our method, where we focus only on the expectation $E[C_{n,m}]$ to achieve our lower bound, does not succeed in the case $m = 2$ is explained by the lack of the term h_n on the right hand side of (1.6). Theorem 1.3 is shown using the following exact expression for the expected running time of Quickselect; see also Section 6 of [9].

Theorem 1.4 (Knuth [13])**.** *Let $C_{n,m}$ be the number of comparisons made by Quickselect to locate the m^{th} smallest of n distinct numbers. Then for all $n \ge m \ge 1$*

$$E[C_{n,m}] = 2[n + 3 + (n+1)h_n - (m+2)h_m - (n-m+3)h_{n-m+1}]. \tag{1.4}$$

In particular,

$$\begin{aligned} E[C_{n,1}] &= 2n - 2h_n && (1.5)\\ E[C_{n,2}] &= 2n - 4 + \frac{2}{n} && (1.6)\\ E[C_{n,3}] &= 2n - \frac{25}{3} + 2h_n + \frac{2}{n-1} \quad \text{and} \\ E[C_{n,4}] &= 2n - 13 + 4h_n - \frac{2}{n} + \frac{2}{n-2}. \end{aligned}$$

Theorems 1.1 and 1.2 are derived by applying Theorem 1.5 that quantifies the if direction of the fixed point property (1.2) in the Wasserstein, or d_1 metric between two random variables X and Y, given by

$$d_1(X,Y) = \sup_{h\in \mathrm{Lip}_1} |Eh(X) - Eh(Y)| \quad \text{where} \quad \mathrm{Lip}_1 = \{h : |h(y) - h(x)| \le |y - x|\}. \tag{1.7}$$

On the left hand side of (1.7) we have chosen to write $d_1(X,Y)$, rather than the technically correct expression $d_1(\mathcal{L}(X), \mathcal{L}(Y))$, only for notational convenience.

Theorem 1.5. *Let W be a non-negative random variable with finite mean, let $\theta > 0$, and let the law of W^* be given by* (1.1)*. Then*

$$d_1(W, D_\theta) \le (1+\theta) d_1(W^*, W).$$

As the Wasserstein distance also satisfies

$$d_1(X,Y) = \inf E|X - Y| \tag{1.8}$$

where the infimum is over all couplings (X, Y) having the given marginals, as is achieved here (see [18], for instance), Theorem 1.5 implies that

$$d_1(W, D_\theta) \le (1+\theta)E|W^* - W| \tag{1.9}$$

for any non-negative random variable W with finite mean, and W^* defined on a common space having the $\mathcal{D}_\theta$-bias distribution of W.

In Section 2 we detail the workings of the Quickselect algorithm and prove Theorems 1.1 and 1.2 by applying Theorem 1.5, which is proved in Section 3. The proof of Theorem 1.3 appears in Section 4.

In related work, [8] considers the Quicksort method, which produces a fully sorted list, and [5] obtains distributional bounds for the running time of a variation of Quickselect to a non-Dickman approximand; compare its characterizion in (1.4) there to (1.1) here.

2 The Quickselect method and the proofs of Theorems 1.1 and 1.2

In this section we apply Theorem 1.5 to obtain the bounds in Theorems 1.1 and 1.2 on the error of the Dickman approximation for the distribution of $W_{n,m}$ in (1.3), the properly normalized running time of the Quickselect algorithm for finding the m^{th} smallest element of a list of n distinct numbers. When the value of m is clear from context, we will write C_n for $C_{n,m}$.

2.1 Quickselect: the case $m = 1$

In this section we prove Theorem 1.1 for the distribution of the number C_n of comparisons that Quickselect requires to locate the smallest element of a list of n distinct numbers. Clearly, a list of size zero requires no comparisons, hence $C_0 = 0$. For $n \ge 1$, the procedure requires the $n-1$ comparisons of the pivot to every other element at the first stage, followed by the cost of processing the left subtree, which may be empty. Since the pivot is chosen uniformly, we obtain the stochastic recursion

$$C_n = n - 1 + C_{V_1} \quad \text{for } n \ge 1\text{, with boundary condition } C_0 = 0, \tag{2.1}$$

where V_1, the size of the left subtree, is a discrete uniform variable on $\{0, \dots, n-1\}$. From (2.1) we see that $C_1 = 0$ and $C_2 = 1$ a.s., and that non-trivial distributions arise for $n \ge 3$.

Before proceeding to the proof of the theorem we describe how for all $n \ge 1$ we may write C_n as a function $C(n; \mathbf{U}_1)$ with

$$\mathbf{U}_k = (U_k, U_{k+1}, \dots) \quad \text{for } k \ge 1,$$

and $U_1, U_2, \dots$ a sequence of i.i.d. uniform variables on $[0, 1]$. Consider the initial list of size $V_0 = n$ as making up the left subtree at stage 0. At stage $k \ge 1$, given a non-null left subtree from the previous stage of size V_{k-1}, a new left subtree of size

$$V_k = \lfloor V_{k-1} U_k \rfloor \quad \text{for } k \ge 1 \tag{2.2}$$

results by choosing a pivot uniformly from the current left subtree. In particular, the conditional distribution of V_k given V_{k-1} satisfies $V_k \sim \mathcal{U}\{0, \dots, V_{k-1} - 1\}$. Rewriting (2.1) in this notation we have

$$C(n; \mathbf{U}_1) = n - 1 + C(\lfloor nU_1 \rfloor; \mathbf{U}_2) \quad \text{for } n \ge 1\text{, with} \quad C(0; \mathbf{U}_k) = 0 \quad \text{for all } k \ge 1. \tag{2.3}$$

As the size of each non-null left subtree decrements by at least one at each iteration, the value of C_n will only depend on an initial subsequence of $\mathbf{U}_1$ of length at most n.

We pause to prove a lemma that is needed in this and the following section.

Lemma 2.1. *If for c a non-negative number and q a positive integer*

$$e_n \le c + \frac{1}{n}\sum_{u=q}^{n-1} e_u \quad \textit{for all } n \ge q, \tag{2.4}$$

then

$$e_n \le c\log(en/q) \quad \textit{for } n \ge q. \tag{2.5}$$

Proof. As (2.4) holds for $n = q$ we see that $e_q \le c$, verifying that the inequality in (2.5) holds at q. Assuming inequality (2.4) holds for $q \le u \le n-1$ for some $n \ge q+1$ we have

$$\begin{aligned} e_n &\le c + \frac{1}{n}\sum_{u=q}^{n-1} e_u \le c + \frac{c}{n}\sum_{u=q}^{n-1}\log(eu/q) \le c + \frac{c}{n}\int_q^n \log(eu/q)du \\ &= c\left(1 + \frac{1}{n}\left[u\log(eu/q) - u\right]\Big|_q^n\right) = c\left(1 + \frac{1}{n}\left[n\log(en/q) - n\right]\right) = c\log(en/q), \end{aligned}$$

completing the inductive step, and the proof. □

We now prove Theorem 1.1. In the proof, we use Lemmas 2.2 and 2.4, which appear with their proofs at end of this section.

Proof of Theorem 1.1. Take $n \ge 1$. With V_k as in (2.2), by (2.3) the variable W_n as given by (1.3) satisfies

$$W_n = \frac{1}{n}C(n;\mathbf{U}_1) - 1 = \frac{1}{n}(n - 1 + C(V_1;\mathbf{U}_2)) - 1 = \frac{1}{n}(C(V_1;\mathbf{U}_2) - 1).$$

We now construct a variable with the W_n^* distribution by first constructing W_n' having the W_n distribution. As $\mathbf{U}_1$ and $\mathbf{U}_2$ are equidistributed,

$$W_n' := \frac{1}{n}C(n,\mathbf{U}_2) - 1 =_d \frac{1}{n}C(n,\mathbf{U}_1) - 1 = W_n,$$

and hence

$$W_n^* := U_1(W_n' + 1) = \frac{1}{n}U_1C(n;\mathbf{U}_2)$$

has the $\mathcal{D}$-bias distribution by (1.1). The difference

$$W_n^* - W_n = \frac{1}{n}\left(U_1C(n;\mathbf{U}_2) - C(V_1;\mathbf{U}_2) + 1\right)$$

satisfies

$$nE|W_n^* - W_n| \le e_n + 1, \quad \text{where we set} \quad e_k = E|U_1C(k;\mathbf{U}_2) - C(\lfloor kU_1\rfloor;\mathbf{U}_2)|, \quad k \ge 0,$$

hence consequence (1.9) of Theorem 1.5 with $\theta = 1$ yields

$$d_1(W_n, D) \le 2E|W_n^* - W_n| \le \frac{2}{n}(e_n + 1). \tag{2.6}$$

We claim that

$$\begin{aligned} e_n &= E|U_1C(n;\mathbf{U}_2) - C(\lfloor nU_1\rfloor;\mathbf{U}_2)| \\ &\le E|U_1(n-1) - \lfloor nU_1\rfloor + 1| + E|U_1C(\lfloor nU_2\rfloor;\mathbf{U}_3) - C(\lfloor\lfloor nU_1\rfloor U_2\rfloor;\mathbf{U}_3)|. \end{aligned}$$

When $\lfloor nU_1 \rfloor \geq 1$ this inequality follows from using the basic recursion (2.3) on both terms forming the difference that defines e_n, followed by applying the triangle inequality, and is easily verified to hold directly in the case $\lfloor nU_1 \rfloor = 0$ by applying (2.3) only on the first term of that difference, noting the second one in this case is zero. Now using that $|u(n-1) - \lfloor nu \rfloor + 1| \leq 2$ for all $u \in [0,1]$, we obtain

$$\begin{aligned}
e_n &\leq 2 + E|U_1 C(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \\
&\leq 2 + E|U_1 C(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3)| + E|C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \\
&= 2 + Ee_{\lfloor nU_2 \rfloor} + E|C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \\
&\leq 4 + Ee_{\lfloor nU_2 \rfloor}. \qquad (2.7)
\end{aligned}$$

For the final term, the inequality

$$E|C(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \leq 2 \quad \text{for all } n \geq 0$$

follows by applying Lemma 2.2, below, that shows that $|\lfloor U_1 \lfloor nU_2 \rfloor \rfloor - \lfloor U_2 \lfloor nU_1 \rfloor \rfloor| \leq 1$ a.s, and Lemma 2.4, also below, that shows that $E|C(p, \mathbf{U}_3) - C(p-1, \mathbf{U}_3)| \leq 2$ for all $p \geq 1$.

Expanding the expectation in $Ee_{\lfloor nU_2 \rfloor}$ in (2.7), using the fact that $\lfloor nU_2 \rfloor$ is uniformly distributed over $\{0, \ldots, n-1\}$ and that $e_0 = e_1 = 0$ by virtue of $C_0 = C_1 = 0$, we obtain

$$e_n \leq 4 + \frac{1}{n} \sum_{u=0}^{n-1} e_u \leq 4 + \frac{1}{n} \sum_{u=2}^{n-1} e_u \quad \text{for } n \geq 2.$$

As $e_1 = 0$ inequality (2.6) shows that the claim of the theorem holds for $n = 1$. Applying Lemma 2.1 with $c = 4$ and $q = 2$ shows that $e_n \leq 4 \log(en/2)$ for $n \geq 2$, and substituting this bound into (2.6) and simplifying now completes the proof. $\square$

We now prove Lemmas 2.2 and 2.4.

Lemma 2.2. *For all* $(u_1, u_2) \in [0,1)^2$ *and* $n \geq 0$,

$$|\lfloor u_1 \lfloor nu_2 \rfloor \rfloor - \lfloor u_2 \lfloor nu_1 \rfloor \rfloor| \leq 1.$$

Proof. Consider the case $n \geq 1$, as otherwise the claim is trivial. Let $s = \lfloor nu_1 \rfloor$ and $t = \lfloor nu_2 \rfloor$, so that $(s,t) \in \{0, 1, \ldots, n-1\}^2$ and

$$s \leq nu_1 < (s+1) \quad \text{and} \quad t \leq nu_2 < (t+1).$$

Then

$$\frac{st}{n} \leq u_2 \lfloor nu_1 \rfloor < \frac{s(t+1)}{n} \quad \text{and} \quad \frac{st}{n} \leq u_1 \lfloor nu_2 \rfloor < \frac{(s+1)t}{n}.$$

Taking the difference,

$$|u_1 \lfloor nu_2 \rfloor - u_2 \lfloor nu_1 \rfloor| < \frac{1}{n} \max\{s, t\} < 1.$$

As the difference between $u_1 \lfloor nu_2 \rfloor$ and $u_2 \lfloor nu_1 \rfloor$ is less than 1, their integer parts can differ by at most 1. $\square$

To prove Lemma 2.4, we will use the easily verified fact that

$$0 \leq \frac{k-1}{p-1} < \frac{k}{p} < \frac{k}{p-1} \leq 1 \quad \text{for } p \geq 2 \text{ and } 1 \leq k \leq p-1, \qquad (2.8)$$

and for $u \in [0,1]$ that

$$(\lfloor (p-1)u \rfloor, \lfloor pu \rfloor) = \begin{cases} (k-1, k-1) & u \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right) \\ (k-1, k) & u \in \left[\frac{k}{p}, \frac{k}{p-1}\right). \end{cases} \qquad (2.9)$$

We will also require the following inequality that can be shown directly using induction.

Lemma 2.3. *If* $c \geq 0, f_1 = 0$ *and*

$$f_p \leq c + \frac{1}{p(p-1)} \sum_{k=1}^{p-1} k f_k \quad \text{for all } p \geq 2$$

then $f_p \leq 2c$ *for all* $p \geq 1$.

Lemma 2.4. *For all* $p \geq 1$

$$f_p := E|C(p, \boldsymbol{U}_1) - C(p-1, \boldsymbol{U}_1)| \leq 2.$$

Proof. As $f_1 = 0$ we need only consider $p \geq 2$. In view of (2.8) we may write

$$\begin{aligned} f_p &= E|C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1)| \\ &= \sum_{k=1}^{p-1} E\left[|C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1)| \,\middle|\, U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right) \right] P\left(U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right)\right) \\ &\quad + \sum_{k=1}^{p-1} E\left[|C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1)| \,\middle|\, U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right) \right] P\left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right)\right). \end{aligned}$$

We claim that the conditional expectation in the first sum is 1. Indeed, for the given range of U_1 the first case of (2.9) yields $(\lfloor (p-1)U_1 \rfloor, \lfloor pU_1 \rfloor) = (k-1, k-1)$, and now (2.3) implies that on this event

$$C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1) = p - 1 + C(k-1, \mathbf{U}_2) - (p - 2 + C(k-1, \mathbf{U}_2)) = 1.$$

For the second sum, the second case of (2.9) yields $(\lfloor (p-1)U_1 \rfloor, \lfloor pU_1 \rfloor) = (k-1, k)$, and

$$\begin{aligned} C(p, \mathbf{U}_1) - C(p-1, \mathbf{U}_1) &= p - 1 + C(k, \mathbf{U}_2) - (p - 2 + C(k-1, \mathbf{U}_2)) \\ &= 1 + C(k, \mathbf{U}_2) - C(k-1, \mathbf{U}_2). \end{aligned}$$

Hence,

$$\begin{aligned} f_p &= \sum_{k=1}^{p-1} P\left(U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right)\right) \\ &\quad + \sum_{k=1}^{p-1} E\left[|1 + C(k, \mathbf{U}_2) - C(k-1, \mathbf{U}_2)| \,\middle|\, U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right) \right] P\left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right)\right) \\ &\leq \sum_{k=1}^{p-1} P\left(U_1 \in \left[\frac{k-1}{p-1}, \frac{k}{p}\right)\right) + \sum_{k=1}^{p-1} (1 + f_k)\, P\left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right)\right) \\ &= 1 + \sum_{k=1}^{p-1} f_k P\left(U_1 \in \left[\frac{k}{p}, \frac{k}{p-1}\right)\right) = 1 + \frac{1}{p(p-1)} \sum_{k=1}^{p-1} k f_k. \end{aligned}$$

Invoking Lemma 2.3 with $c = 1$ now completes the proof. □

2.2 Case of $m \geq 2$

In this section we prove Theorem 1.2 for the approximation of the distribution of the properly scaled value of the number $C_{n,m}$ of comparisons made by the Quickselect algorithm Q_m to determine the m^{th} smallest element of a list of n distinct numbers in the case $m \geq 2$.

As the m^{th} smallest element of the list does not exist when $n < m$, no comparisons are required and we may set $C_{n,m} = 0$ over this range. In the non-trivial case $n \geq m$, Q_m

begins as for $m=1$ at the first stage by selecting a uniformly chosen pivot, giving rise, through $n-1$ comparisons to the pivot, to a left subtree of size V_1, uniformly distributed over $\{0,\ldots,n-1\}$, and a right subtree of size $n-1-V_1$. If $V_1 \geq m$ then the m^{th} smallest element of the original list lies in the left subtree, and we may locate it by applying Q_m to it. If $V_1 = m-1$ then the pivot is the m^{th} smallest element and the process stops. Otherwise $V_1 < m-1$, and the m^{th} smallest element is the $m-V_1-1^{st}$ smallest element in the right subtree, which we then locate by applying Q_{m-V_1-1} to it. Hence, we obtain

$$\begin{aligned}&C_{n,m} = 0 \quad \text{for } 0 \leq n \leq m-1, \text{ and}\\ &C_{n,m} = n-1+C_{V_1,m}\mathbf{1}(V_1 \geq m)+C_{n-V_1-1,m-V_1-1}\mathbf{1}(V_1 < m-1) \quad \text{for } n \geq m. \qquad (2.10)\end{aligned}$$

We now develop a simple bound on the expectation $E[C_{n,m}]$.

Lemma 2.5. *Let $C_{n,m}$ be the number of Quickselect comparisons for locating the m^{th} smallest element of a list of n distinct numbers. Then for all $m \geq 1$,*

$$E[C_{n,m}] \leq 4n \quad \textit{for all } n \geq 0.$$

Proof. Recall h_n is the harmonic series $\sum_{1\leq k\leq n} 1/k$ for $n \geq 1$. The claim is trivial unless $n \geq m$, and is also easily seen to be true for $m=1$ and $m=2$ using (1.5) and (1.6). Hence, we take $n \geq m \geq 3$.

For such n and m, writing the difference between the two harmonic series below as a sum and separating out the last term for $j=m-2$, we have

$$(n-m+3)(h_n-h_{n-m+1}) = \sum_{j=0}^{m-3}\frac{n-m+3}{n-j}+1+\frac{1}{n-m+2} \leq m, \qquad (2.11)$$

the inequality holding since each ratio is bounded by 1. Hence, using the expression given for $E[C_{n,m}]$ in Theorem 1.4 and applying (2.11) to yield the first inequality below, we obtain the upper bound

$$\begin{aligned}E[C_{n,m}] &= 2[n+3+(n+1)h_n-(m+2)h_m-(n-m+3)h_{n-m+1}]\\ &= 2[n+3+(m-2)h_n-(m+2)h_m+(n-m+3)(h_n-h_{n-m+1})]\\ &\leq 2[n+3+(m-2)h_n-(m+2)h_m+m]\\ &= 2[n+3+(m+1)(h_n-h_m)-3h_n+m-h_m]\\ &\leq 2[n+(m+1)\left(\frac{n-m}{m+1}\right)-3(h_n-1)+m-h_m]\\ &= 2[2n-3(h_n-1)-h_m] \leq 4n. \qquad \square\end{aligned}$$

Note that the indicator on the first term on the right hand side of (2.10) may be dropped, due to the boundary condition there, on the line above. Now letting $C_m(n;\mathbf{U}_1)$ be defined by rewriting (2.10) as (2.3) was derived from (2.1), we obtain

$$\begin{aligned}&C_m(n;\mathbf{U}_1) = 0 \quad \text{for } 0 \leq n \leq m-1, \text{ and otherwise}\\ &\qquad C_m(n;\mathbf{U}_1) = n-1+C_m(\lfloor nU_1\rfloor;\mathbf{U}_2)\\ &\qquad\quad + C_{m-1-\lfloor nU_1\rfloor}(n-1-\lfloor nU_1\rfloor;\mathbf{U}_2)\mathbf{1}(\lfloor nU_1\rfloor < m-1). \qquad (2.12)\end{aligned}$$

We next provide the following result that parallels Lemma 2.4 for the case $m=1$.

Lemma 2.6. *For all $m \geq 2$ and $p \geq 1$*

$$f_p := E|C_m(p;\boldsymbol{U}_1)-C_m(p-1;\boldsymbol{U}_1)| \leq 2+16m.$$

Proof. As $C_m(p;\mathbf{U}_1)=0$ for all $0\le p\le m-1$ we may take $p\ge m$. By the basic recursion (2.12) we have

$$\begin{aligned}
C_m(p;\mathbf{U}_1)-C_m(p-1;\mathbf{U}_1) &= 1+C_m(\lfloor pU_1\rfloor;\mathbf{U}_2)-C_m(\lfloor (p-1)U_1\rfloor;\mathbf{U}_2)\\
&\quad + C_{m-1-\lfloor pU_1\rfloor}(p-1-\lfloor pU_1\rfloor;\mathbf{U}_2)\mathbf{1}(\lfloor pU_1\rfloor<m-1)\\
&\quad - C_{m-1-\lfloor (p-1)U_1\rfloor}(p-2-\lfloor (p-1)U_1\rfloor;\mathbf{U}_2)\mathbf{1}(\lfloor (p-1)U_1\rfloor<m-1)\\
&:= 1+(C_m(\lfloor pU_1\rfloor;\mathbf{U}_2)-C_m(\lfloor (p-1)U_1\rfloor;\mathbf{U}_2))+R.
\end{aligned}$$

Applying the triangle inequality and taking expectation yields

$$f_p\le 1+E|C_m(\lfloor pU_1\rfloor;\mathbf{U}_2)-C_m(\lfloor (p-1)U_1\rfloor;\mathbf{U}_2)|+E|R|. \tag{2.13}$$

For the first expectation in (2.13), by (2.9) we have

$$\begin{aligned}
&E|C_m(\lfloor pU_1\rfloor;\mathbf{U}_2)-C_m(\lfloor (p-1)U_1\rfloor;\mathbf{U}_2)|\\
&=\sum_{k=1}^{p-1}E\left[|C_m(k-1,\mathbf{U}_2)-C_m(k-1,\mathbf{U}_2)|\,\middle|\,U_1\in\left[\frac{k-1}{p-1},\frac{k}{p}\right)\right]P\left(U_1\in\left[\frac{k-1}{p-1},\frac{k}{p}\right)\right)\\
&\quad+\sum_{k=1}^{p-1}E\left[|C_m(k,\mathbf{U}_2)-C_m(k-1,\mathbf{U}_2)|\,\middle|\,U_1\in\left[\frac{k}{p},\frac{k}{p-1}\right)\right]P\left(U_1\in\left[\frac{k}{p},\frac{k}{p-1}\right)\right)\\
&=\sum_{k=1}^{p-1}f_kP\left(U_1\in\left[\frac{k}{p},\frac{k}{p-1}\right)\right)=\frac{1}{p(p-1)}\sum_{k=1}^{p-1}kf_k.
\end{aligned}$$

Now applying Lemma 2.5 on the first term of the remainder R, and using that $\lfloor pU_1\rfloor\sim\mathcal{U}\{0,\ldots,p-1\}$, yields

$$\begin{aligned}
E[C_{m-1-\lfloor pU_1\rfloor}(p-1-\lfloor pU_1\rfloor;\mathbf{U}_2)\mathbf{1}(\lfloor pU_1\rfloor<m-1)] &\le \frac{4}{p}\sum_{k=0}^{m-2}(p-1-k)\\
&\le \frac{4}{p}(p-1)(m-1)\le 4m,
\end{aligned}$$

and replacing p by $p-1$ we see that the same bound holds for the expectation of the final term of R.

Substituting the bounds achieved into (2.13) we obtain

$$f_p\le 1+8m+\frac{1}{p(p-1)}\sum_{k=1}^{p-1}kf_k \quad \text{for all } p\ge m. \tag{2.14}$$

As $f_p=0$ for $1\le p\le m-1$ inequality (2.14) holds for all $p\ge 2$, and the conditions for invoking Lemma 2.3 with $c=1+8m$ are satisfied, yielding the desired conclusion. □

Proof of Theorem 1.2. Let $n\ge m$. From (1.3) and (2.12), letting $V_1=\lfloor nU_1\rfloor$,

$$\begin{aligned}
W_n &= \frac{1}{n}C_m(n;\mathbf{U}_1)-1\\
&= \frac{1}{n}(n-1+C_m(V_1;\mathbf{U}_2)+C_{m-1-V_1}(n-1-V_1;\mathbf{U}_2)\mathbf{1}(V_1<m-1))-1\\
&= \frac{1}{n}(C_m(V_1;\mathbf{U}_2)+C_{m-1-V_1}(n-1-V_1;\mathbf{U}_2)\mathbf{1}(V_1<m-1)-1).
\end{aligned} \tag{2.15}$$

We now construct a variable with the W_n^* distribution. As $\mathbf{U}_1$ and $\mathbf{U}_2$ are equidistributed, W_n' given by the first equality in (2.15) when substituting $\mathbf{U}_2$ in place of $\mathbf{U}_1$ has law $\mathcal{L}(W_n)$. Hence, by (1.1) with $\theta = 1$, letting

$$W_n^* = U_1(W_n' + 1) = \frac{1}{n} U_1 C_m(n; \mathbf{U}_2), \tag{2.16}$$

the pair (W_n, W_n^*) is a coupling of a variable with the W_n distribution to one with its Dickman $\mathcal{D}$-bias distribution. Applying consequence (1.9) of Theorem 1.5, we obtain

$$d_1(W_n, D) \le \frac{2}{n} f_n \quad \text{where} \quad f_n = nE|W_n^* - W_n|. \tag{2.17}$$

Letting

$$e_n = E|U_1 C_m(n; \mathbf{U}_2) - C_m(\lfloor nU_1 \rfloor; \mathbf{U}_2)|,$$

in view of (2.15) and (2.16), and applying Lemma 2.5 to bound expectations of the form $E[C_{n,m}]$ and that $V_1 \sim \mathcal{U}\{0, 1, \ldots, n-1\}$, we obtain

$$\begin{aligned}
f_n &= nE|W_n^* - W_n| \\
&= E|U_1 C_m(n; \mathbf{U}_2) - C_m(V_1; \mathbf{U}_2) - C_{m-1-V_1}(n-1-V_1; \mathbf{U}_2)\mathbf{1}(V_1 < m-1) + 1| \\
&\le e_n + E|C_{m-1-V_1}(n-1-V_1; \mathbf{U}_2)\mathbf{1}(V_1 < m-1)| + 1 \\
&\le e_n + \frac{4}{n}\sum_{k=0}^{m-2}(n-1-k) + 1 \\
&\le e_n + \frac{4}{n}(n-1)(m-1) + 1 \le e_n + 4m.
\end{aligned} \tag{2.18}$$

To control e_n, invoke the basic recursion (2.12) to write

$$\begin{aligned}
U_1 C_m(n; \mathbf{U}_2) &= U_1(n-1) + U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) \\
&\quad + U_1 C_{m-1-\lfloor nU_2 \rfloor}(n-1-\lfloor nU_2 \rfloor; \mathbf{U}_3)\mathbf{1}(\lfloor nU_2 \rfloor < m-1) \\
&= U_1(n-1) + U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) + R_1
\end{aligned}$$

where

$$R_1 = U_1 C_{m-1-\lfloor nU_2 \rfloor}(n-1-\lfloor nU_2 \rfloor; \mathbf{U}_3)\mathbf{1}(\lfloor nU_2 \rfloor < m-1),$$

and similarly,

$$\begin{aligned}
C_m(\lfloor nU_1 \rfloor; \mathbf{U}_2) &= (\lfloor nU_1 \rfloor - 1)\mathbf{1}(\lfloor nU_1 \rfloor \ge m) + C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) + R_2 \\
&= (\lfloor nU_1 \rfloor - 1) + C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) + R_2 + R_3
\end{aligned}$$

where

$$R_2 = C_{m-1-\lfloor \lfloor nU_1 \rfloor U_2 \rfloor}(\lfloor nU_1 \rfloor - 1 - \lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)\mathbf{1}(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor < m-1, \lfloor nU_1 \rfloor \ge m),$$

and

$$R_3 = -(\lfloor nU_1 \rfloor - 1)\mathbf{1}(\lfloor nU_1 \rfloor \le m-1).$$

Taking the expectation of the absolute difference and using that $|u(n-1) - \lfloor nu \rfloor + 1| \le 2$ for all $u \in [0,1]$, we obtain

$$\begin{aligned} e_n = E|U_1 C_m(n; \mathbf{U}_2) - C_m(\lfloor nU_1 \rfloor; \mathbf{U}_2)| \\ \le E|U_1(n-1) - (\lfloor nU_1 \rfloor - 1)| + E|U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) \\ - C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| + E|R_1| + E|R_2| + E|R_3| \\ \le 2 + E|U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| + E|R_1| + E|R_2| + E|R_3| \\ \le 2 + E|U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C_m(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3)| \\ + E|C_m(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| + E|R_1| + E|R_2| + E|R_3|. \end{aligned} \tag{2.19}$$

Lemmas 2.2 and 2.6 yield

$$E|C_m(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3) - C_m(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3)| \le 2 + 16m. \tag{2.20}$$

For the first remainder term R_1, by Lemma 2.5, we have

$$\begin{aligned} E|R_1| = \frac{1}{2} E[C_{m-1-\lfloor nU_2 \rfloor}(n - 1 - \lfloor nU_2 \rfloor; \mathbf{U}_3) \mathbf{1}(\lfloor nU_2 \rfloor \le m-2)] \\ \le \frac{2}{n} \sum_{k=0}^{m-2} (n-1-k) = \frac{2}{n}(n-1)(m-1) \le 2m. \end{aligned} \tag{2.21}$$

For R_2, we first condition on the event $\lfloor nU_1 \rfloor = k$, where the presence of the indicator $\lfloor nU_1 \rfloor \ge m$ restricts the range to $m \le k \le n-1$, then further on $\lfloor kU_2 \rfloor = j$, where the indicator $\mathbf{1}(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor < m-1)]$ imposes $0 \le j \le m-2$. The values of k are taken uniformly over $\{0, \ldots, n-1\}$, and, given k, those of j uniformly over $\{0, \ldots, k-1\}$. Applying Lemma 2.5 then yields

$$\begin{aligned} E|R_2| = E[C_{m-1-\lfloor \lfloor nU_1 \rfloor U_2 \rfloor}(\lfloor nU_1 \rfloor - 1 - \lfloor \lfloor nU_1 \rfloor U_2 \rfloor; \mathbf{U}_3) \mathbf{1}(\lfloor \lfloor nU_1 \rfloor U_2 \rfloor < m-1, \lfloor nU_1 \rfloor \ge m)] \\ \le \frac{4}{n} \sum_{k=m}^{n-1} \frac{1}{k} \sum_{j=0}^{m-2} (k-1-j) \le \frac{4m}{n} \sum_{k=1}^{n-1} \frac{1}{k}(k-1) \le \frac{4m}{n}(n-1) \le 4m. \end{aligned} \tag{2.22}$$

As R_3 satisfies

$$E|R_3| = E|(\lfloor nU_1 \rfloor - 1) \mathbf{1}(\lfloor nU_1 \rfloor \le m-1)| \le m, \tag{2.23}$$

substituting the bounds (2.20)-(2.23) into (2.19) yields that, for all $n \ge m$,

$$\begin{aligned} e_n \le 4 + 23m + E|U_1 C_m(\lfloor nU_2 \rfloor; \mathbf{U}_3) - C_m(\lfloor \lfloor nU_2 \rfloor U_1 \rfloor; \mathbf{U}_3)| \\ = 4 + 23m + \frac{1}{n} \sum_{k=0}^{n-1} e_k = 4 + 23m + \frac{1}{n} \sum_{k=m}^{n-1} e_k, \end{aligned}$$

where the final equality follows by noting that $C(k; \mathbf{U}_1) = 0$ for $k \le m-1$. Applying Lemma 2.1 yields that, for all $n \ge m$,

$$e_n \le (4 + 23m) \log(ne/m),$$

and now from (2.18) we conclude

$$f_n \le e_n + 4m = (4 + 23m) \log(ne/m) + 4m.$$

Substitution into (2.17), and simplification, yields the claim. □

3 Proof of Theorem 1.5

Theorem 1.5 was originally proven using Stein's method in [10], but [14] offered the following much simpler approach.

Proof. Let $U \sim \mathcal{U}[0,1]$ be independent of the pair (W, D_θ), which are constructed on the same space so as to achieve the infimum in (1.8). Then, as $D_\theta =_d D_\theta^*$,

$$d_1(W^*, D_\theta) = d_1(U^{1/\theta}(W+1), U^{1/\theta}(D_\theta+1)) \le E[U^{1/\theta}|W - D_\theta|] = \frac{\theta}{\theta+1} d_1(W, D_\theta).$$

Now, by the triangle inequality,

$$d_1(W, D_\theta) \le d_1(W, W^*) + d_1(W^*, D_\theta) \le d_1(W, W^*) + \frac{\theta}{\theta+1} d_1(W, D_\theta).$$

Rearranging the inequality yields the claimed bound. □

4 Proof of Theorem 1.3

We now apply Theorem 1.4 to prove Theorem 1.3.

Proof of Theorem 1.3. Since $f(x) = x$ is an element of Lip_1, expression (1.7) for the Wasserstein distance yields that

$$d_1(W_{n,m}, D) \ge |E[W_{n,m}] - E[D]| = |E[W_{n,m}] - 1| = \left| \frac{1}{n} E[C_{n,m}] - 2 \right|,$$

applying (1.3) and that (see e.g. [12]) $E[D] = 1$.

Now, slightly rewriting the equality in (1.4) as

$$E[C_{n,m}] = 2\left[n + 3 + (m-2)h_n - (m+2)h_m + (n-m+3)(h_n - h_{n-m+1})\right]$$

for $m > 2$ we have

$$\frac{1}{n} E[C_{n,m}] - 2 \ge \frac{2[(m-2)h_n - (m+2)h_m + 3]}{n} \ge \frac{2[(m-2)\log n - |(m+2)h_m - 3|]}{n},$$

using $h_n > \log n$. Hence, the claim of Theorem 1.3 holds for $m > 2$. We see the claim of Theorem also holds for $m = 1$ by using the form (1.5), which yields $|E[C_{n,m}/n - 2| = 2h_n/n$, noting that in this case $(m+2)h_m - 3 = 0$. □

References

[1] Richard Arratia, A. D. Barbour, and Simon Tavaré, *Logarithmic combinatorial structures: a probabilistic approach*, EMS Monographs in Mathematics, European Mathematical Society (EMS), Zürich, 2003. MR-2032426

[2] Ehsan Azmoodeh, Benjamin Arras, Guillaume Poly, and Yvik Swan, *Distances between probability distributions via characteristic functions and biasing*, arXiv preprint:1605.06819 (2016).

[3] A. D. Barbour and Bruno Nietlispach, *Approximation by the Dickman distribution and quasi-logarithmic combinatorial structures*, Electron. J. Probab. **16** (2011), no. 29, 880–902. MR-2793242

[4] Chinmoy Bhattacharjee and Larry Goldstein, *Dickman approximation in simulation, summations and perpetuities*, to appear in: Bernoulli, arXiv preprint:1706.08192 (2018).

[5] Benjamin Dadoun and Ralph Neininger, *A statistical view on exchanges in Quickselect*, ANALCO14—Meeting on Analytic Algorithmics and Combinatorics, SIAM, Philadelphia, PA, 2014, pp. 40–51. MR-3248352

[6] Luc Devroye and Omar Fawzi, *Simulating the Dickman distribution*, Statist. Probab. Lett. **80** (2010), no. 3-4, 242–247. MR-2575452

[7] Karl Dickman, *On the frequency of numbers containing prime factors of a certain relative magnitude*, Arkiv for matematik, astronomi och fysik **22** (1930), no. 10, 1–14.

[8] James Allen Fill and Svante Janson, *Quicksort asymptotics*, J. Algorithms **44** (2002), no. 1, 4–28, Analysis of algorithms. MR-1932675

[9] James Allen Fill and Takéhiko Nakama, *Analysis of the expected number of bit comparisons required by Quickselect*, Algorithmica **58** (2010), no. 3, 730–769. MR-2672478

[10] Larry Goldstein, *Non asymptotic distributional bounds for the Dickman approximation of the running time of the Quickselect algorithm*, arXiv preprint:1703.00505v1 (2017).

[11] Charles AR Hoare, *Algorithm 65: find*, Communications of the ACM **4** (1961), no. 7, 321–322.

[12] Hsien-Kuei Hwang and Tsung-Hsi Tsai, *Quickselect and the Dickman function*, Combin. Probab. Comput. **11** (2002), no. 4, 353–371. MR-1918722

[13] Donald E. Knuth, *Mathematical analysis of algorithms, in Information processing 71 (Proc. IFIP Congress, Ljubljana, 1971), Vol. 1: Foundations and systems*, (1972), 19–27. MR-0403310

[14] Ralph Neininger, *Private communication*, (2017).

[15] Mathew D. Penrose and Andrew R. Wade, *Random minimal directed spanning trees and Dickman-type distributions*, Adv. in Appl. Probab. **36** (2004), no. 3, 691–714. MR-2079909

[16] Ross G. Pinsky, *On the strange domain of attraction to generalized Dickman distribution for sums of independent random variables*, Electron. J. Probab. **23** (2018), Paper No. 3, 17. MR-3751078

[17] Ross G Pinsky, *A natural probabilistic model on the integers and its relation to Dickman-type distributions and Buchstab's function*, to appear in: Probability and Analysis in Interacting Physical Systems - in Honor of S.R.S. Varadhan", Springer, arXiv preprint:1606.02965 (2019).

[18] Svetlozar T. Rachev, *Probability metrics and the stability of stochastic models*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Ltd., Chichester, 1991. MR-1105086

Acknowledgments. The author thanks Ralph Neininger for his vast simplification of the previous proof of Theorem 1.5 in the preprint [10], as well as for the suggestion for obtaining the lower bounds as achieved in Theorem 1.3. The author also sincerely thanks two reviewers whose suggestions and observations were extremely valuable, which included pointing out that Theorem 1.4 is a known result due to Knuth, and a simplification of Lemma 2.5.

Electronic Journal of Probability
Electronic Communications in Probability

Advantages of publishing in EJP-ECP

- Very high standards
- Free for authors, free for readers
- Quick publication (no backlog)
- Secure publication (LOCKSS[1])
- Easy interface (EJMS[2])

Economical model of EJP-ECP

- Non profit, sponsored by IMS[3], BS[4] , ProjectEuclid[5]
- Purely electronic

Help keep the journal free and vigorous

- Donate to the IMS open access fund[6] (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

[1]LOCKSS: Lots of Copies Keep Stuff Safe http://www.lockss.org/
[2]EJMS: Electronic Journal Management System http://www.vtex.lt/en/ejms.html
[3]IMS: Institute of Mathematical Statistics http://www.imstat.org/
[4]BS: Bernoulli Society http://www.bernoulli-society.org/
[5]Project Euclid: https://projecteuclid.org/
[6]IMS Open Access Fund: http://www.imstat.org/publications/open.htm