

On the tails of the limiting Quicksort distribution*

Svante Janson[†]

Abstract

We give asymptotics for the left and right tails of the limiting Quicksort distribution. The results agree with, but are less precise than, earlier non-rigorous results by Knessl and Spankowski.

Keywords: Quicksort; binary search tree; internal pathlength; tail asymptotics.

AMS MSC 2010: Primary 68P10, Secondary 60C05; 60E05.

Submitted to ECP on August 31, 2015, final version accepted on November 2, 2015.

Supersedes arXiv:1508.07230v2.

1 Introduction

Let X_n be the number of comparisons used by the algorithm Quicksort when sorting n distinct numbers, initially in a uniformly random order. Equivalently, X_n is the internal pathlength in a random binary search tree with n nodes. (See e.g. Knuth [7, Sections 5.2.2 and 6.2.2] or Drmota [1, Chapter 8 and Section 1.4.1] for a description of the algorithm and of binary search trees.) It follows that X_n satisfies the distributional recurrence relation

$$X_n \stackrel{d}{=} X_{U_n-1} + X_{n-U_n}^* + n - 1, \quad n \geq 1, \quad (1.1)$$

where $\stackrel{d}{=}$ denotes equality in distribution, and, on the right, U_n is distributed uniformly on the set $\{1, \dots, n\}$, $X_j^* \stackrel{d}{=} X_j$, $X_0 = 0$, and $U_n, X_0, \dots, X_{n-1}, X_0^*, \dots, X_{n-1}^*$ are all independent. (Thus, (1.1) can be regarded as a definition of X_n .)

It is well-known, and easy to show from (1.1), that

$$\mathbb{E} X_n = 2(n+1)H_n - 4n \sim 2n \ln n, \quad (1.2)$$

where $H_n := \sum_{k=1}^n k^{-1}$ is the n :th harmonic number. Moreover, it was proved by Régnier [9] and Rösler [10], using different methods, that the normalized variables

$$Z_n := \frac{X_n - \mathbb{E} X_n}{n} \quad (1.3)$$

converge in distribution to some limiting random variable Z , as $n \rightarrow \infty$.

There is no simple description of the distribution of Z , but various results have been shown by several different authors. For example, Z has an everywhere finite moment generating function, and thus all moments are finite [10], with $\mathbb{E} Z = 0$ and $\text{Var} Z = 7 - \frac{2}{3}\pi^2$; furthermore, Z has a density which is infinitely differentiable [11; 2].

*Partly supported by the Knut and Alice Wallenberg Foundation

[†]Department of Mathematics, Uppsala University, Uppsala, Sweden. E-mail: svante.janson@math.uu.se

Moreover, the recurrence relation (1.1) yields in the limit a distributional identity, which can be written as

$$Z \stackrel{d}{=} UZ' + (1 - U)Z'' + g(U), \tag{1.4}$$

where U , Z' and Z'' are independent, $U \sim U(0, 1)$ is uniform, $Z', Z'' \stackrel{d}{=} Z$, and g is the deterministic function

$$g(u) := 2u \ln u + 2(1 - u) \ln(1 - u) + 1. \tag{1.5}$$

Furthermore, Rösler [10] showed that (1.4) together with $\mathbb{E}Z = 0$ and $\text{Var}Z < \infty$ determines the distribution of Z uniquely; see further [3]. The identity (1.4) is the basis of much of the study of Z , including the present work.

In the present paper we study the asymptotics of the tail probabilities $\mathbb{P}(Z \leq -x)$ and $\mathbb{P}(Z \geq x)$ as $x \rightarrow \infty$. Using non-rigorous methods from applied mathematics (assuming an as yet unverified regularity hypothesis), Knessl and Szpankowski [6] found very precise asymptotics of both the left tail and the right tail. Their result for the left tail is that, as $x \rightarrow \infty$, with $\gamma = (2 - \frac{1}{\ln 2})^{-1}$,

$$\mathbb{P}(Z \leq -x) = (c_1 + o(1)) \exp(-c_2 e^{\gamma x}) = \exp(-e^{\gamma x + c_3 + o(1)}), \tag{1.6}$$

where c_1, c_2, c_3 are some constants (c_1 is explicit in [6], but not c_2). For the right tail, they give a more complicated expression, which by ignoring higher order terms implies, for example,

$$\mathbb{P}(Z \geq x) = \exp(-x \ln x - x \ln \ln x + (1 + \ln 2)x + o(x)). \tag{1.7}$$

It has been a challenge to justify these asymptotics rigorously, and so far very little progress has been made. Some rigorous upper bounds were given by Fill and Janson [4], in particular

$$\mathbb{P}(Z \geq x) \leq \exp(-x \ln x + (1 + \ln 2)x), \quad x \geq 303, \tag{1.8}$$

with the same leading term (in the exponent) as (1.7), and for the left tail

$$\mathbb{P}(Z \leq -x) \leq \exp(-x^2/5), \quad x \geq 0, \tag{1.9}$$

which is much weaker than (1.6).

Also the present paper falls short of the (non-rigorous) asymptotics (1.6)–(1.7) from [6], but we show, by simple methods, the following results, which at least show that the leading terms in the top exponents in (1.6)–(1.7) are correct.

Theorem 1.1. (i) Let $\gamma := (2 - \frac{1}{\ln 2})^{-1}$. As $x \rightarrow \infty$,

$$\exp(-e^{\gamma x + \ln \ln x + O(1)}) \leq \mathbb{P}(Z \leq -x) \leq \exp(-e^{\gamma x + O(1)}) \tag{1.10}$$

(ii) As $x \rightarrow \infty$,

$$\exp(-x \ln x - x \ln \ln x + O(x)) \leq \mathbb{P}(Z \geq x) \leq \exp(-x \ln x + O(x)). \tag{1.11}$$

We show the lower bounds in Sections 3 and 4, and the upper bounds in Sections 5 and 6. The lower bounds are proved by direct arguments using the identity (1.4); the upper bounds are proved by the standard method of first estimating the moment generating function.

Remark 1.2. The right inequality in (1.11) follows from the more precise (1.8), where an explicit value is given for the implicit constant; we include this part of (1.11) for completeness. (The proof in Section 6 actually yields a better constant than (1.8) for large x , see (6.10).) We expect that, similarly, the implicit constants in the other parts of (1.10)–(1.11) could be replaced by explicit bounds, using more careful versions of the arguments and estimates below. However, in order to keep the proofs simple, we have not attempted this.

Remark 1.3. We consider only the limiting random variable Z , and not Z_n or X_n for finite n . Of course, the results for Z imply corresponding results for the tails $\mathbb{P}(Z_n \leq -x)$ and $\mathbb{P}(Z_n \geq x)$ for n sufficiently large (depending on x), but we do not attempt to give any explicit results for finite n . For some bounds for finite n , see [5] and (for large deviations) [8].

Remark 1.4. Although we do not work with Z_n for finite n , the proofs below of the lower bounds can be interpreted for finite n , saying that we can obtain $Z_n \leq -x$ with roughly the given probability (for large n) by considering the event that in the first $\Theta(x)$ generations, all splits are close to balanced (with proportions $\frac{1}{2} \pm x^{-1/2}$, say); similarly, to obtain $Z_n \geq x$ we let there be one branch of length $\Theta(x)$ where all splits are extremely unbalanced (with at most a fraction $(x \ln x)^{-1}$ on the other side). The fact that we require an exponential number of splits to be extreme for the lower tail, but only a linear number for the right tail, can be seen as an explanation of the difference between the two tails, with the left tail doubly exponential and the right tail roughly exponential.

2 Preliminaries

Note that g in (1.5) is a continuous convex function on $[0, 1]$, with maximum $g(0) = g(1) = 1$ and minimum $g(1/2) = 1 - 2 \ln 2 = -(2 \ln 2 - 1) < 0$.

Let $\psi(t) := \mathbb{E} e^{tZ}$ be the moment generating function of Z . As said above, Rösler [10] showed that $\psi(t)$ is finite for every real t . The distributional identity (1.4) yields, by conditioning on U , the functional equation

$$\psi(t) := \mathbb{E} e^{tZ} = \int_0^1 \psi(ut)\psi((1-u)t)e^{tg(u)} du. \tag{2.1}$$

We may replace Z by the right-hand side of (1.4); hence we may without loss of generality assume the equality (not just in distribution)

$$Z = UZ' + (1-U)Z'' + g(U). \tag{2.2}$$

3 Left tail, lower bound

Proof of lower bound in (1.10). Let $\varepsilon > 0$ be so small that $g(\frac{1}{2} + \varepsilon) < 0$, and let $a := -g(\frac{1}{2} + \varepsilon) > 0$. For any z , on the event $\{Z' \leq -z, Z'' \leq -z, \text{ and } |U - \frac{1}{2}| \leq \varepsilon\}$, (2.2) yields

$$Z \leq -Uz - (1-U)z + g(U) = -z + g(U) \leq -z - a. \tag{3.1}$$

Hence, for any real z ,

$$\mathbb{P}(Z \leq -z - a) \geq 2\varepsilon \mathbb{P}(Z \leq -z)^2. \tag{3.2}$$

It follows by induction that

$$\mathbb{P}(Z \leq -na) \geq (2\varepsilon)^{2^n - 1} \mathbb{P}(Z \leq 0)^{2^n}, \quad n \geq 0. \tag{3.3}$$

Consequently, using $2\varepsilon \leq 1$, $\mathbb{P}(Z \leq -na) \geq (2\varepsilon \mathbb{P}(Z \leq 0))^{2^n}$, and thus, with $c := \ln(2 \mathbb{P}(Z \leq 0)) > -\infty$,

$$\ln \mathbb{P}(Z \leq -na) \geq 2^n (\ln \varepsilon + c), \quad n \geq 0. \tag{3.4}$$

If $x > 0$, we take $n = \lceil x/a \rceil$ and obtain

$$\ln \mathbb{P}(Z \leq -x) \geq 2^{x/a+1}(\ln \varepsilon + c). \quad (3.5)$$

We choose (for large x) $\varepsilon = x^{-1/2}$, so, using Taylor's formula,

$$a = -g\left(\frac{1}{2} + \varepsilon\right) = -g\left(\frac{1}{2}\right) + O(\varepsilon^2) = 2 \ln 2 - 1 + O(x^{-1}) \quad (3.6)$$

and thus

$$a^{-1} = (2 \ln 2 - 1)^{-1} + O(x^{-1}). \quad (3.7)$$

Consequently, (3.5) yields

$$\ln \mathbb{P}(Z \leq -x) \geq 2^{x/(2 \ln 2 - 1) + O(1)}(\ln x^{-1/2} + c) = -e^{\gamma x + O(1) + \ln \ln x}. \quad (3.8)$$

□

4 Right tail, lower bound

Proof of lower bound in (1.11). Let $0 < \delta < \frac{1}{2}$. If $0 < U \leq \delta$, then

$$g(U) \geq g(\delta) = 1 + 2\delta \ln \delta + O(\delta) \geq 1 + 3\delta \ln \delta, \quad (4.1)$$

with the last inequality holding provided δ is small enough.

Assume that (4.1) holds, and assume that $Z' \geq 0$, $Z'' \geq z \geq 0$ and $U \leq \delta$. Then (2.2) yields

$$Z \geq (1 - \delta)z + g(\delta) \geq z - \delta z + 1 - 3\delta \ln \delta^{-1}. \quad (4.2)$$

Consequently,

$$\mathbb{P}(Z \geq z + 1 - \delta z - 3\delta \ln \delta^{-1}) \geq \delta \mathbb{P}(Z \geq 0) \mathbb{P}(Z \geq z). \quad (4.3)$$

Let x be sufficiently large and choose $\delta = 1/(x \ln x)$. Then, for $0 \leq z \leq x$,

$$z + 1 - \delta z - 3\delta \ln \delta^{-1} \geq z + 1 - \frac{1}{\ln x} - 3 \frac{\ln(x \ln x)}{x \ln x} \geq z + 1 - \frac{2}{\ln x}, \quad (4.4)$$

provided x is large enough. Hence, if $b := 1 - \frac{2}{\ln x}$ and $c := \mathbb{P}(Z \geq 0) > 0$, then for $0 \leq z \leq x$ we have

$$\mathbb{P}(Z \geq z + b) \geq c\delta \mathbb{P}(Z \geq z). \quad (4.5)$$

By induction, we find for $0 \leq n \leq x/b + 1$,

$$\mathbb{P}(Z \geq nb) \geq c^n \delta^n \mathbb{P}(Z \geq 0) = c^{n+1} \delta^n > (c\delta)^{n+1}. \quad (4.6)$$

Consequently, taking $n := \lceil x/b \rceil$,

$$\begin{aligned} \ln \mathbb{P}(Z \geq x) &\geq (n+1)(\ln c + \ln \delta) \geq (x/b + 2)(\ln c + \ln \delta) \\ &= (x + O(x/\ln x))(-\ln x - \ln \ln x + O(1)) \\ &= -x \ln x - x \ln \ln x + O(x). \end{aligned} \quad (4.7)$$

□

5 Left tail, upper bound

Lemma 5.1. *There exists $a \geq 0$ such that for all $t > 0$, with $\kappa := \gamma^{-1} = 2 - \frac{1}{\ln 2}$,*

$$\psi(-t) < \exp(\kappa t \ln t + at + 1). \tag{5.1}$$

Proof. We note that $t \ln t \geq -e^{-1}$ for $t > 0$, and thus $\kappa t \ln t + at + 1 \geq -\kappa e^{-1} + 1 > 0$. Since $\psi(t)$ is continuous and $\psi(0) = 1$, there exists $t_1 > 0$ such that $\psi(-t) < \exp(1 - \kappa e^{-1})$ for $0 \leq t \leq t_1$, and thus (5.1) holds for all such t , and any $a \geq 0$. Next, let $t_2 := \pi e^2$. We may choose $a > 0$ such that (5.1) holds for $t \in [t_1, t_2]$.

Before proceeding to larger t , define

$$h(u) := u \ln u + (1 - u) \ln(1 - u) \tag{5.2}$$

and note that $g(u) = 2h(u) + 1$ by (1.5).

Now suppose that (5.1) fails for some $t > 0$ and let $T := \inf\{t > 0 : (5.1) \text{ fails}\}$. Then $T \geq t_2$, and, by continuity,

$$\psi(-T) = \exp(\kappa T \ln T + aT + 1). \tag{5.3}$$

Furthermore, if $0 < u < 1$, then (5.1) holds for $t = uT$ and $t = (1 - u)T$, and thus, recalling (5.2),

$$\begin{aligned} \psi(-uT)\psi(-(1-u)T) &< \exp(\kappa uT \ln(uT) + \kappa(1-u)T \ln((1-u)T) + auT + a(1-u)T + 2) \\ &= \exp(\kappa T \ln T + \kappa(u \ln u + (1-u) \ln(1-u))T + aT + 2) \\ &= \exp(\kappa T \ln T + \kappa h(u)T + aT + 2). \end{aligned}$$

Furthermore, $g(u) = 1 + 2h(u)$, and thus we obtain

$$\psi(-uT)\psi(-(1-u)T)e^{-Tg(u)} \leq \exp(\kappa T \ln T - ((2 - \kappa)h(u) + 1)T + aT + 2). \tag{5.4}$$

By (5.2), $h(u)$ is a convex function with $h(\frac{1}{2}) = -\ln 2$, $h'(\frac{1}{2}) = 0$ and $h''(u) = u^{-1} + (1-u)^{-1} \geq 4$, and thus by Taylor's formula, $h(u) \geq -\ln 2 + 2(u - \frac{1}{2})^2$. Furthermore, $2 - \kappa = 1/\ln 2$, and thus

$$(2 - \kappa)h(u) + 1 \geq \frac{2}{\ln 2}(u - \frac{1}{2})^2 \geq (u - \frac{1}{2})^2. \tag{5.5}$$

Combining (2.1), (5.4), and (5.5), we obtain

$$\begin{aligned} \psi(-T) &\leq \int_0^1 \exp(\kappa T \ln T + aT + 2 - (u - \frac{1}{2})^2 T) du \\ &< \exp(\kappa T \ln T + aT + 2) \int_{-\infty}^{\infty} e^{-(u - \frac{1}{2})^2 T} du \\ &= \sqrt{\frac{\pi}{T}} \exp(\kappa T \ln T + aT + 2). \end{aligned} \tag{5.6}$$

Since $T \geq t_2 = \pi e^2$, this yields $\psi(-T) < \exp(\kappa T \ln T + aT + 1)$, which contradicts (5.3). This contradiction shows that no such T exists, and thus (5.1) holds for all $t > 0$. \square

Proof of upper bound in (1.10). For $x \geq 0$ and any $t \geq 0$, by Lemma 5.1,

$$\mathbb{P}(Z \leq -x) \leq e^{-tx} \mathbb{E} e^{-tZ} = e^{-tx} \psi(-t) < \exp(-tx + \kappa t \ln t + at + 1). \tag{5.7}$$

We optimize by taking $t = \exp(\kappa^{-1}(x - a) - 1)$ and obtain

$$\ln \mathbb{P}(Z \leq -x) < t(\kappa \ln t + a - x) + 1 = -\kappa t + 1 = -e^{\kappa^{-1}x + O(1)}, \tag{5.8}$$

which is the upper bound in (1.10) because $\kappa^{-1} = \gamma$. \square

6 Right tail, upper bound

As said in the introduction, (1.8) was proved in [4]. Nevertheless we give for completeness a proof of the upper bound in (1.11), similar to the proof in Section 5. (It is also similar to the proof in [4] but simpler, partly because we do not keep track of all constants and do not try to optimize; nevertheless, it yields a slight improvement of (1.8) for large x , see (6.10) below.)

Lemma 6.1. *There exists $a \geq 0$ such that for all $t \geq 0$,*

$$\psi(t) \leq \exp(e^t + at). \tag{6.1}$$

Note that [4, Corollary 4.3] shows the bound $\psi(t) \leq \exp(2e^t)$ for $t \geq 5.02$, which is explicit, but weaker for large t .

Proof. Since $\psi(0) = 1 < e$, it follows by continuity that there exists $t_1 > 0$ such that $\psi(t) \leq e$ for $t \in [0, t_1]$, and thus (6.1) holds for $t \in [0, t_1]$ and any $a \geq 0$.

Let $t_2 := 100$, and choose a so that (6.1) holds for $t \in [t_1, t_2]$. Assume that (6.1) fails for some $t > 0$, and let $T := \inf\{t > 0 : (6.1) \text{ fails}\}$. Then $T \geq t_2$, and, by continuity,

$$\psi(T) = \exp(e^T + aT). \tag{6.2}$$

Furthermore, if $0 < u < 1$, then (6.1) holds for $t = uT$ and $t = (1 - u)T$, and thus, using (2.1) and the symmetry $u \leftrightarrow 1 - u$ there, and $g(u) \leq 1$,

$$\begin{aligned} \psi(T) &\leq 2 \int_0^{1/2} \exp\left(e^{uT} + auT + e^{(1-u)T} + a(1-u)T + Tg(u)\right) du \\ &\leq 2 \int_0^{1/2} \exp\left(e^{uT} + e^{T-uT} + aT + T\right) du. \end{aligned} \tag{6.3}$$

We consider two cases.

(i) If $uT \leq 1$, then $e^{-uT} \leq 1 - \frac{1}{2}uT$, and thus

$$e^{uT} + e^{T-uT} + aT + T \leq e + e^T\left(1 - \frac{1}{2}uT\right) + (a + 1)T. \tag{6.4}$$

Hence, the contribution to (6.3) for $u \leq 1/T$ is no more than

$$\begin{aligned} 2 \int_0^{1/T} \exp\left(e^T + (a + 1)T + e - \frac{1}{2}Te^T u\right) du &< 2 \exp\left(e^T + (a + 1)T + e\right) \frac{1}{\frac{1}{2}Te^T} \\ &= \frac{4e^e}{T} \exp(e^T + aT) \leq 0.7\psi(T), \end{aligned} \tag{6.5}$$

by (6.2) and $T \geq t_2 = 100$, since $4e^e \doteq 60.62$.

(ii) For $uT > 1$ and $u < \frac{1}{2}$, recalling $T \geq t_2 = 100$,

$$\begin{aligned} e^{uT} + e^{T-uT} + aT + T &\leq 2e^{T-uT} + aT + T \leq 2e^{-1}e^T + aT + T \\ &\leq 0.8e^T + T + aT \leq 0.9e^T + aT \\ &= e^T + aT - 0.1e^T \leq e^T + aT - 100. \end{aligned} \tag{6.6}$$

Hence, the contribution to (6.3) for $uT > 1$ is less than, recalling (6.2),

$$\exp(e^T + aT - 100) = e^{-100}\psi(T) < 0.1\psi(T). \tag{6.7}$$

Using (6.5) and (6.7) in (6.3), we find

$$\psi(T) < 0.7\psi(T) + 0.1\psi(T), \tag{6.8}$$

a contradiction. Hence T cannot exist and (6.1) holds for all $t \geq 0$. \square

Proof of upper bound in (1.11). For $x \geq 0$ and any $t \geq 0$, by Lemma 6.1,

$$\mathbb{P}(Z \geq x) \leq e^{-tx} \mathbb{E} e^{tZ} = e^{-tx} \psi(t) \leq \exp(-tx + e^t + at). \quad (6.9)$$

We take $t = \ln x$ (assuming $x \geq 1$) and obtain

$$\mathbb{P}(Z \geq x) \leq \exp(-x \ln x + x + O(\ln x)), \quad x \geq 1. \quad (6.10)$$

(The optimal choice of t is actually $\ln(x - a)$, but this leads to the same result up to $o(1)$ in the exponent, which is absorbed by the error term $O(\ln x)$.) \square

Acknowledgments. I thank David Belius and Jim Fill for helpful comments.

References

- [1] Michael Drmota, *Random Trees*. Springer, Vienna, 2009. MR-2484382
- [2] James Allen Fill and Svante Janson, Smoothness and decay properties of the limiting Quicksort density function. *Mathematics and Computer Science (Proceedings, Colloquium on Mathematics and Computer Science, Versailles 2000)*, eds. D. Gardy and A. Mokkadem, Birkhäuser, Basel, 2000, pp. 53–64. MR-1798287
- [3] James Allen Fill and Svante Janson, A characterization of the set of fixed points of the Quicksort transformation. *Electronic Comm. Probab.* **5** (2000), no. 9, 77–84. MR-1781841
- [4] James Allen Fill and Svante Janson, Approximating the limiting Quicksort distribution. *Random Structures Algorithms* **19** (2001), no. 3-4, 376–406. MR-1871560
- [5] James Allen Fill and Svante Janson, Quicksort asymptotics. *J. Algorithms* **44** (2002), no. 1, 4–28. MR-1932675
- [6] Charles Knessl and Wojciech Szpankowski, Quicksort algorithm again revisited. *Discrete Math. Theor. Comput. Sci.* **3** (1999), 43–64. MR-1695194
- [7] Donald E. Knuth, *The Art of Computer Programming. Vol. 3: Sorting and Searching*. 2nd ed., Addison-Wesley, Reading, Mass., 1998. MR-0378456
- [8] C. J. H. McDiarmid and R. B. Hayward, Large deviations for Quicksort. *J. Algorithms* **21** (1996), no. 3, 476–507. MR-1417660
- [9] Mireille Régnier, A limiting distribution for quicksort. *RAIRO Inform. Théor. Appl.* **23** (1989), no. 3, 335–343. MR-1020478
- [10] Uwe Rösler, A limit theorem for “Quicksort”. *RAIRO Inform. Théor. Appl.* **25** (1991), no. 1, 85–100. MR-1104413
- [11] Kok Hooi Tan and Petros Hadjicostas, Some properties of a limiting distribution in Quicksort. *Statist. Probab. Lett.* **25** (1995), 87–94. MR-1364822

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)

Economical model of EJP-ECP

- Low cost, based on free software (OJS¹)
- Non profit, sponsored by IMS², BS³, PKP⁴
- Purely electronic and secure (LOCKSS⁵)

Help keep the journal free and vigorous

- Donate to the IMS open access fund⁶ (click here to donate!)
- Submit your best articles to EJP-ECP
- Choose EJP-ECP over for-profit journals

¹OJS: Open Journal Systems <http://pkp.sfu.ca/ojs/>

²IMS: Institute of Mathematical Statistics <http://www.imstat.org/>

³BS: Bernoulli Society <http://www.bernoulli-society.org/>

⁴PK: Public Knowledge Project <http://pkp.sfu.ca/>

⁵LOCKSS: Lots of Copies Keep Stuff Safe <http://www.lockss.org/>

⁶IMS Open Access Fund: <http://www.imstat.org/publications/open.htm>