Average case analysis of the merging algorithm of Hwang and Lin

W. Fernandez de la Vega* A.M.Frieze† M. Santha*

Abstract

We derive an asymptotic equivalent to the average running time of the merging algorithm of Hwang and Lin applied on two linearly ordered lists of numbers $a_1 < a_2 ... < a_m$ and $b_1 < b_2 ... < b_n$ when m and n tend to infinity in such a way that the ratio $\rho = \frac{m}{n}$ is constant. We show that the distribution of the running time is concentrated around its expectation except when ρ is a power of 2. When ρ is a power of 2, we obtain an asymptotic equivalent for the expectation of the running time.

Key words: merging, average case analysis.

1 Introduction

In the merging problem we are given two linearly ordered lists of numbers $a_1 < a_2 < ... < a_m$ and $b_1 < b_2 < ... < b_n$ and the task consists of pooling these two lists into a third ordered list $c_1 < c_2 ... < c_{n+m}$. We assume that the $n+m$ elements are distinct and $n \leq m$. The merging is performed by pairwise comparisons between the elements in the two lists. The measure of complexity of a merging algorithm is the number of comparisons made by the algorithm, and the complexity of the merging problem is the complexity of the best merging algorithm. As usual, we can speak of worst case and average case complexity.

The worst case complexity of the merging problem is quite well studied. The most widely used merging algorithm which performs well for any values of m and n is *binary merge* was invented by Hwang and Lin [5]. They have shown that the complexity of binary merge in the worst case is at most $L(n,m) + n$, where $L(n,m) = \log_2 \binom{n+m}{n}$ is the obvious lower bound coming from information theory. Several much more complicated merging algorithms were proposed later by Christen [1] and Manacher [7] which perform better than binary merge in the worst case for some ratios m/n, but binary merge is still the best deterministic algorithm when $m/n \leq 3$. Recently Fernandez de la Vega, Kannan and Santha [3] have proposed a probabilistic merging algorithm which is always faster than binary merge for $m/n > (\sqrt{5}+1)/2 \approx 1.618$.

In contrast with the worst case complexity, very little is known about the average case complexity of merging. However, Fernandez de la Vega, Kannan and Santha [3] have derived a lower bound for the class of *insertive* merging algorithms, in which for each element of the smaller list, the comparisons involving it are made consecutively. They proved that if $(1+\eta) \leq m/n \leq (\sqrt{2}+1-\eta)$

*CNRS, URA 410, LRI, Université Paris-Sud, ,91405 Orsay, France.
†Department of Mathematical Sciences, Carnegie Mellon University. Supported in part by NSF grant CCR-9530974.

for some constant $\eta > 0$, then every insertive merging algorithm makes on the average at least a constant factor more comparisons than the information theoretical lower bound.

The purpose of this paper is to derive estimates for the average running time of the merging algorithm of Hwang and Lin. This algorithm (or rather a simplified version of it, sufficient for our purposes) will be described in the next section where we also give a precise definition of the probability model.

2 The merging algorithm of Hwang and Lin

Let $a_1 < a_2 < ... < a_m$ and $b_1 < b_2 < ... < b_n$ denote two ordered lists of numbers, where $m \geq n$. Note that merging these two lists is equivalent to determining the numbers $J(i)$, for $1 \leq i \leq n$, where $J(i)$ denotes the index of the rightmost term of the first sequence which is smaller than b_i. Plainly, $J(i)$ is the number of terms of the first sequence which are smaller than b_i. This means that $a_{J(i)} < b_i < a_{J(i)+1}$ if $a_1 < b_i < a_m$, $J(i) = m$ if $b_i > a_m$ and we set $J(i) = 0$ if $b_i < a_1$.

Roughly speaking, the merging algorithm of Hwang and Lin proceeds (as virtually all known merging algorithms) by succesive insertions of the elements of one of the given lists into the other. One peculiarity of this algorithm is that the roles of the lists may be interchanged during the computation, precisely at any moment when the relative order between the sizes of the parts of the lists remaining to be merged is reversed. However, it can be shown that, for any fixed ratio $\frac{m}{n} = \rho > 1$ of the sizes of the two lists and $n \to \infty$, with probability $1 - o(1)$, most of the time during the execution of the algorithm, the size of the part of the first list remaining to be merged will stay bigger than that of the second. Thus, the algorithm of Hwang and Lin will keep inserting elements of the second list into the first one except perhaps for the very last steps. Accordingly, we shall analyse a simpler algorithm (with the specification that all the insertions are insertions of elements of B into A), which we call algorithm HL; and check our assertion about the sizes of the unmerged parts of the lists. We refer the reader to the paper of Hwang and Lin ([5]) for the description of their algorithm. For convenience, we shall express the algorithm HL in terms of the $J(i)$'s as follows.

Algorithm HL
$J(0) := 0$
for $i = 1$ **to** n **do**
  $r := \lfloor \log_2 \left(\frac{m-J(i-1)}{n-i+1} \right) \rfloor$
  $s := 2^r$
  $l := \lfloor \frac{m-J(i-1)}{s} \rfloor$
  for $k = 1$ **to** l **do**
    $j := J(i-1) + ks$
    if $b_i < a_j$ **then**
      determine $J(i)$ using binary search in the interval $[j-s, j-1]$;
      continue outer loop
  endfor
  determine $J(i)$ using binary search in the interval $[j+1, m]$
endfor
end

We shall refer to the comparisons used by the binary search as comparisons of the second type. The other comparisons will be called comparisons of the first type.

Throughout the analysis we will make the following hypotheses. The algorithm receives as input pairwise distinct elements, and $n \leq m$. The inputs are randomly chosen with uniform distribution among the $\binom{m+n}{n}$ distinct linear orders which are consistent with the partial orders defined by each of the two lists. The integers m and n tend to infinity, but their proportion remains fixed, that is there exists a constant $\rho > 1$ such that $\frac{m}{n} = \rho$. We shall set $\frac{m}{m+n} = \gamma$ and $\alpha = 1 - \gamma$. We shall prove the following theorems.

Theorem 1 *Let ρ be a fixed constant > 1 and distinct from a power of 2. Let $\gamma = \frac{\rho}{\rho+1}$, $r = \lfloor \log_2 \rho \rfloor$ and $s = 2^r$. The number of comparisons $M(m,n)$ used by the algorithm of Hwang and Lin applied to two lists of length m and n satisfies*

$$\frac{M(m,n)}{n} \ \to \ r + \frac{1}{1-\gamma^s}$$

in probability as m and n tend to infinity with $\frac{m}{n} = \rho$.

Theorem 2 *Let again $M(m,n)$ denote the number of comparisons $M(m,n)$ used by the algorithm of Hwang and Lin applied to two lists of length m and n. Let m and n tend to infinity with $\frac{m}{n} = 2^r$ where r is a fixed positive integer and let $s = 2^{r-1}$. Then the number of comparisons $M(m,n)$ used by the algorithm of Hwang and Lin applied to two lists of length m and n satisfies*

$$\frac{\mathbf{E}M(m,n)}{n} \ \to \ r - \frac{1}{2} + \frac{1}{2(1-\gamma^s)} + \frac{1}{2(1-\gamma^{2s})}$$

as m and $n \to \infty$.

Actually, our results are slightly more precise. The proofs occupy the rest of the paper. Let us first remind the reader of the following well known fact which we state as a lemma for ease of reference.

Lemma 1 *For any integer $r \geq 1$ binary search in an interval of length 2^r requires exactly r comparisons.*

3 Proof of Theorem 1

Let $a_1 < a_2 < ... < a_m$ and $b_1 < b_2 < ... < b_n$ be our two ordered lists and recall that $J(i)$ was defined as the index of the rightmost term of the first list which is smaller than b_i. Equivalently, we can think of $J(i) + i$ as the rank of the ith 1 which appears in a 0-1 string of length $m + n$ chosen uniformly at random from all 0-1 strings with n 1's and m 0's, $m = \rho n$ where $\rho > 1$. Let X_i denote the number of 0's between the ith 1 and the i+1th 1. Thus,

$$J(i) = X_1 + X_2 + \cdots + X_i.$$

We replace X_i by Y_i where $Y_1, Y_2, \ldots, Y_{n+1}$ are independent copies of the geometric random variable Y with parameter $p = \frac{m}{m+n+1} = \gamma + O(1/n)$ where $\gamma = \rho/(1+\rho)$ i.e. $\mathbf{Pr}(Y = j) = p^j(1-p)$ for $j = 0, 1, 2, \ldots \infty$. The next lemma shows that in order to get back to our model, we need only condition on $Y_1 + Y_2 + \cdots + Y_{n+1} = m$.

Lemma 2 *Conditional on $Y_1 + Y_2 + \cdots + Y_{n+1} = m$, each 0-1 string with n 1's and m 0's is equally likely.*

Proof Fix $y_1, y_2, \ldots, y_{n+1}$ such that $y_1 + y_2 + \cdots + y_{n+1} = m$.

$$\begin{aligned} \mathbf{Pr}(Y_i = y_i,\, 1 \le i \le n+1) &= \prod_{i=1}^{n+1} p^{y_i}(1-p) \\ &= p^m (1-p)^{n+1} \end{aligned}$$

is independent of $y_1, y_2, \ldots, y_{n+1}$. □

Lemma 3

$$\mathbf{Pr}(Y_1 + Y_2 + \cdots + Y_{n+1} = m) \ge \frac{1+o(1)}{\sqrt{2\pi\rho(\rho+1)n}}.$$

Proof

$$\begin{aligned} \mathbf{Pr}(Y_1 + Y_2 + \cdots + Y_{n+1} = m) &= \sum_{y_1+y_2+\cdots+y_{n+1}=m} p^m(1-p)^{n+1} \\ &= \binom{n+m}{n} p^m (1-p)^{n+1} \\ &= (1+o(1))\sqrt{\frac{m+n}{2\pi mn}}\frac{(m+n)^{m+n}}{m^m n^n} p^m (1-p)^{n+1} \\ &= \frac{1+o(1)}{\sqrt{2\pi\rho(\rho+1)n}}. \end{aligned}$$

□

Lemma 4 *If $Z_1, Z_2, \ldots, Z_k$ are independent geometric random variables with parameter α then there exists $\epsilon_0 = \epsilon_0(\alpha)$ such that if $\epsilon \le \epsilon_0$ then*

$$\mathbf{Pr}\left(\left|Z_1 + Z_2 + \cdots + Z_k - \frac{k\alpha}{1-\alpha}\right| \ge \frac{\epsilon k \alpha}{1-\alpha}\right) \le 2e^{-\epsilon^2 k/8}.$$

Proof The proof is the usual application of the Markov inequality to the moment generating function of the sum. □

The core of the proof relies on the following lemma.

Lemma 5 *There exists $K = K(\rho)$ such that* **whp** *for any $\epsilon > 0$ and for $0 \le i \le n - K\log n$*

(a) $m - J(i) \ge n - i$.

(b) $\lceil \log_2 \left(\frac{m-J(i)}{n-i}\right) \rceil = \lceil \log_2 \rho \rceil$.

Proof (a) follows from (b).

$$\mathbf{Pr}\left(\left|\frac{m-J(i)}{n-i}\right| \ge \epsilon\rho\right) = \mathbf{Pr}\left(\left|\frac{1}{n-i}\sum_{t=i+1}^{n+1} X_t - \rho\right| \ge \epsilon\rho\right)$$

$$
\begin{aligned}
&= \mathbf{Pr}\left(\left|\frac{1}{n-i}\sum_{t=i+1}^{n+1} Y_t - \rho\right| \geq \epsilon\rho \;\mid \sum_{t=1}^{n+1} Y_t = m\right)\\
&= O(n^{1/2})\mathbf{Pr}\left(\left|\frac{1}{n-i}\sum_{t=i+1}^{n+1} Y_t - \rho\right| \geq \epsilon\rho\right)\\
&= O(n^{1/2})e^{-(1-o(1))\epsilon^2(n-i)/8}
\end{aligned}
$$

from Lemma 3. □

Note that the inequality a) of Lemma 5 implies our claim that up to very near from the end of the execution, (at least up to $i = n - K\log n$), **whp** the roles of the two lists are not reversed.

Theorem 1 can now be easily proved.

Let $n_0 = n - K\log n$. Let T_i denote the number of Type 1 comparisons. Let $\hat{T}_i$ refer to the independent model.

$$
\begin{aligned}
\mathbf{Pr}\left(\left|\sum_{i=1}^{n_0} T_i - \frac{n_0}{1-\gamma^s}\right| \geq \frac{\epsilon\gamma^s n_0}{1-\gamma^s}\right) &= \mathbf{Pr}\left(\left|\sum_{i=1}^{n_0} \hat{T}_i - \frac{n_0}{1-\gamma^s}\right| \geq \frac{\epsilon\gamma^s n_0}{1-\gamma^s} \mid \sum_{t=1}^{n+1} Y_t = m\right)\\
&= O(n^{1/2})\mathbf{Pr}\left(\left|\sum_{i=1}^{n_0} \hat{T}_i - \frac{n_0}{1-\gamma^s}\right| \geq \frac{\epsilon\gamma^s n_0}{1-\gamma^s}\right)
\end{aligned}
$$

The algorithm HL tells us that $T_i - 1$ is the number of consecutive intervals of length $s = 2^r$ laying at the right of a_i and which contain only 0's. Thus $\hat{T}_i - 1$ has a geometric distribution with parameter $1-\gamma^s$. Hence,

$$
\mathbf{Pr}\left(\left|\sum_{i=1}^{n_0}(\hat{T}_i - 1) - \frac{n_0\gamma^s}{1-\gamma^s}\right| \geq \frac{\epsilon\gamma^s n_0}{1-\gamma^s}\right) \leq e^{-\epsilon^2 n_0/8}. \tag{1}
$$

To finish the proof of Theorem 1 we only have to argue that $\sum_{i=n_0+1}^{n}\hat{T}_i = O(\log n)$ **whp** which also follows from Lemma 4, and to use Lemma 1.

4 Proof of Theorem 2

Here again we analyse in fact algorithm HL.

Let, for $1 \leq i \leq n$, $Z_i = X_i - i\rho$, $S_i = \sum_{j=1}^{i} Z_j$, and $S_o = 0$,. Let $n_1 = n^{1/2}(\log n)^4$. The cases corresponding to $\rho = 2^r$ bring a new difficulty: In these cases, the parameter s for the $i+1$th step is equal to 2^{r-1} when S_i is positive, and to 2^r when S_i is non-negative.

Let A denote the event

$$
\max_{1\leq i\leq n} |S_i| \leq \sqrt{n}\log n. \tag{2}
$$

Lemma 6 *The event A occurs* **whp**.

Proof (Sketch) Bound above for each i the probability $\mathbf{Pr}[\hat{S}_i > \sqrt{n}\log n]$, where $\hat{S}_i$ refers to the independent model, check that the sum of these bounds is $o(n^{-1/2})$ and use Lemma 2. □

Lemma 6 and the fact that the total number of comparisons is $\Omega(n)$ show that we can condition on A for the proof of Theorem 2.

Lemma 7 *In the range $n_1 \leq i \leq n - n_1$, both probabilities* $\mathbf{Pr}[S_i \leq 0]$ *and* $\mathbf{Pr}[S_i > 0]$ *are asymptotic to* $1/2$.

Proof Use the formula

$$\mathbf{Pr}[S_i \leq 0] = \frac{\mathbf{Pr}[\hat{S}_i \leq 0 \ \bigwedge \ \hat{S}_n = 0]}{\mathbf{Pr}[\hat{S}_n = 0]},$$

where $\hat{S}_i$ refers to the independent model, and use normal approximation. □

Lemma 8 *a) Assume $i \leq n - n_1$ and $0 \leq S_i \leq \sqrt{n}\log n$. Then,*

$$\mathbf{E}(T_i) = \frac{1 + o(1)}{1 - \gamma^{2^r}}$$

b) Assume $i \leq n - n_1$ and $-\sqrt{n}\log n \leq S_i < 0$. Then,

$$\mathbf{E}(T_i) = \frac{1 + o(1)}{1 - \gamma^{2^{r-1}}}.$$

Proof This follows as in (1). Note that we can condition on $S_i \geq 0$, say, and then replace $X_{i+1}, \ldots, X_{n+1}$ by independent $Y_{i+1}, \ldots, Y_{n+1}$ subject to $Y_{i+1} + \cdots + Y_{n+1} = m - (X_1 + \cdots + X_i)$. The parameter for $Y_j, j > i$ will have to change slightly, but will remain close enough to γ **whp** within our range. □

We have of course

$$\mathbf{E}T_i = \mathbf{Pr}[S_i \leq 0]\mathbf{E}[T_i|S_i \leq 0] + \mathbf{Pr}[S_i > 0].\mathbf{E}[T_i|S_i > 0], \ 1 \leq i \leq n.$$

Lemmas 7 and 8 give then, assuming A, the estimate

$$\mathbf{E}T_i = \left(\frac{1}{2} + o(1)\right)\left(\frac{1}{1 - \gamma^s} + \frac{1}{1 - \gamma^{2s}}\right),$$

with $s = 2^{r-1}$, for each i in the interval $n_1 \leq i \leq n - n_1$. In order to conclude the proof of Theorem 2, it remains only, using the linearity of expectation, to account for the Type 2 comparisons, and to check that the contribution of the expectations corresponding to the remaining values of i is $o(n)$.

□

References

[1] C. Christen (1978) *Improving the bound on optimal merging*, Proceedings of 19th FOCS, pp. 259-266.

[2] V. Chvatal (1984) *Probabilistic methods in graph theory*, Annals of Operations Research 1, pp. 171-182.

[3] W. Fernandez de la Vega, S. Kannan and M. Santha (1990) *Two Probabilistic Results on Merging* SIAM J. COMPUT. 22, N2, pp. 261-271, 1993.

[4] F. K. Hwang and S. Lin (1971), *Optimal merging of 2 elements with n elements*, Acta Informatica 1, pp. 145-158.

[5] F. K. Hwang and S. Lin (1972), *A simple algorithm for merging two disjoint linearly ordered lists*, SIAM J. COMPUT. 1, pp. 31-39.

[6] D. E. Knuth (1973), The Art of Computer Programming, Volume 3: Sorting and Searching, Addison-Wesley.

[7] G. K. Manacher (1979), *Significant improvements to the Hwang-Ling merging algorithm*, JACM, Vol. 26, No. 3, pp. 434-440.