Approximate Sorting*

Joachim Giesen†, Eva Schuberth†, and Miloš Stojaković†

† Institute for Theoretical Computer Science, ETH Zürich, CH-8092 Zürich

Abstract. We show that any randomized algorithm to approximate any given ranking of n items within expected Spearman's footrule distance $n^2/\nu(n)$ needs at least $n\,(\min\{\log \nu(n), \log n\} - 6)$ comparisons. This bound is tight up to a constant factor since there exists a deterministic algorithm that shows that $6n(\log \nu(n) + 1)$ comparisons are always sufficient.

Keywords. Sorting, Ranking, Spearman's footrule metric, Kendall's tau metric

1 Introduction

Our motivation to study approximate sorting comes from the following market research application. We want to find out how a respondent ranks a set of products. In order to simulate real buying situations the respondent is presented pairs of products out of which he has to choose one that he prefers, i.e., he has to perform paired comparisons. The respondent's ranking is then reconstructed from the sequence of his choices. That is, a procedure that presents a sequence of product pairs to the respondent in order to obtain the product ranking is nothing else than a comparison based sorting algorithm. We can measure the efficiency of such an algorithm in terms of the number of comparisons needed in order to obtain the ranking. The information theoretic lower bound on sorting [7] states that there is no procedure that can determine a ranking by posing less than $n \log \frac{n}{e}$ paired comparison questions to the respondent, i.e., in general $\Omega(n \log n)$ comparisons are needed. Even for only moderately large n that easily is too much since respondents often get worn out after a certain number of questions and do not answer further questions faithfully anymore. On the other hand, it might be enough to know the respondent's ranking approximately. In this paper we pursue the question of how many comparisons are necessary and sufficient in order to approximately rank n products.

In order to give sense to the term "approximately" we need some metric to compare rankings. Assume that we are dealing with n products. Since a ranking is a permutation of the products, this means that we need a metric on the permutation group S_n. Not all of the metrics, e.g., the Hamming distance

* Partly supported by the Swiss National Science Foundation under the grant "Robust Algorithms for Conjoint Analysis" and by the joint Berlin/Zurich graduate program Combinatorics, Geometry and Computation, financed by ETH Zurich and the German Science Foundation (DFG).

that counts how many products are ranked differently, are meaningful for our application. For example, if in the respondent's ranking one exchanges every second product with its predecessor, then the resulting ranking has maximal Hamming distance to the original one. Nevertheless, this ranking still tells a lot about the respondent's preferences. In marketing applications Kendall's tau metric [4] is frequently used since it seems to capture the intuitive notion of closeness of two rankings and also arises naturally in the statistics of certain random rankings [8].

Our results. Instead of working with Kendall's metric we use Spearman's footrule metric [4] which essentially is equivalent to Kendall's metric, since the two metrics are within a constant factor of each other [4]. The maximal distance between any two rankings of n products in Spearman's footrule metric is less than n^2. We show that in order to obtain a ranking at distance $n^2/\nu(n)$ to the respondent's ranking with any strategy, a respondent has in general to perform at least $n\,(\min\{\log \nu(n), \log n\} - 6)$ comparisons. Moreover, if we allow the strategy to be randomized such that the obtained ranking is at expected distance $n^2/\nu(n)$ to the respondent's ranking, we can show that the same bound on the minimum number of comparisons holds.

On the other hand, there is a deterministic strategy (algorithm), suggested in [2], that shows that $6n(\log \nu(n) + 1)$ comparisons are always sufficient.

Related work. At first glance our work seems related to work done on pre-sorting. In pre-sorting the goal is to pre-process the data such that fewer comparisons are needed afterwards to sort them. For example in [5] it is shown that with $O(1)$ pre-processing one can save $\Theta(n)$ comparisons for Quicksort on average. Pre-processing can be seen as computing a partial order on the data that helps for a given sorting algorithm to reduce the number of necessary comparisons. The structural quantity that determines how many comparisons are needed in general to find the ranking given a partial order is the number of linear extensions of the partial order, i.e., the number of rankings consistent with the partial order. Actually, the logarithm of this number is a lower bound on the number of comparisons needed in general [6]. Here we study another structural measure, namely, the maximum diameter in the Spearman's metric of the set of rankings consistent with a partial order. Our results shows that with $o(n \log n)$ comparisons one can make this diameter asymptotically smaller than the diameter of the set of all rankings. That is not the case for the number of linear extensions which stays in $\Theta(2^{n \log n})$.

Notation. The logarithm log in this paper is assumed to be binary, and by id we denote the identity (increasing) permutation of $[n]$.

2 Algorithm

The idea of the ASort algorithm is to partition the products into a sorted sequence of equal-sized bins such that the elements in each bin have smaller

rank than any element in subsequent bins. This approach was suggested by Chazelle [2] for near-sorting. The output of the algorithm is the sequence of bins. Note that we do not specify the ordering of elements inside each bin, but consider any ranking consistent with the ordering of the bins. We will show that any such ranking approximates the actual ranking of the elements in terms of Spearman's footrule metric

$$D(\pi, \mathrm{id}) = D(\pi) = \sum_{i=1}^{n} |i - \pi(i)|,$$

where $\pi(i)$ is the rank of the element of rank i in an approximate ranking, i.e., $|i - \pi(i)|$ measures deviation of the approximated rank from the actual rank. Note that for any ranking the distance in the Spearman's footrule metric to id is at most $\frac{n^2}{2}$.

Since for every i the value $|i - \pi(i)|$ is bounded by n divided by the number of bins, we see that the approximation quality depends on the number of bins.

The algorithm ASort iteratively performs a number of median searches, each time placing the median into the right position in the ranking. Here the median of n elements is defined to be the element of rank $\lfloor \frac{n+1}{2} \rfloor$.

```
ASort (B : set, m : int)
1  B_01 := B   //   B_ij is the j'th bin in the i'th round
2  for i := 1 to m do
3     for j := 1 to 2^(i-1) do
4        compute the median of B_(i-1)j
5        B_i(2j-1) := {x ∈ B_(i-1)j | x ≤ median}
6        B_i(2j) := {x ∈ B_(i-1)j | x > median}
7     end for
8  end for
9  return B_m1, ..., B_m(2^m)
```

To compute the median in line 4 and to partition the elements in line 5 and 6 we use the deterministic algorithm by Blum et al. [1] that performs at most $5.73n$ comparisons in order to compute the median of n elements and to partition them according to the median. We note that in putting the algorithm ASort to practice one may want to use a different median algorithm, like, e.g., RandomizedSelect [3].

In the following we determine the number of comparisons the algorithm ASort needs on input B with $|B| = n$ in order to guarantee a prescribed approximation error of the actual ranking for any ranking consistent with the ordering of the bins $B_{m1}, \ldots, B_{m(2^m)}$ computed by the algorithm.

Lemma 1. *For every $x \in B_{ij}$, where $0 \le i \le m$ and $1 \le j \le 2^i$, it holds*

$$\sum_{k=1}^{j-1} |B_{ik}| + 1 \le \mathrm{rank}(x) \le \sum_{k=1}^{j} |B_{ik}|.$$

Proof. The lemma can be proven by induction on the number of rounds. By construction, the elements in B_{11} have rank at least 1 and at most $\lfloor \frac{n+1}{2} \rfloor = |B_{11}|$ and the elements in B_{12} have rank at least $\lfloor \frac{n+1}{2} \rfloor + 1 = |B_{11}| + 1$ and at most $n = |B_{11}| + |B_{12}|$.

Now assume that the statement holds after the $(i-1)$'th round. The algorithm partitions every bin $B_{(i-1)j}$ into two bins $B_{i(2j-1)}$ and $B_{i(2j)}$. Again by construction the elements in bin $B_{i(2j-1)}$ have rank at least

$$\sum_{k=1}^{j-1} |B_{(i-1)k}| + 1 = \sum_{k=1}^{j-1} (|B_{i(2k-1)}| + |B_{i(2k)}|) + 1 = \sum_{k=1}^{(2j-1)-1} |B_{ik}| + 1,$$

and at most

$$\sum_{k=1}^{(2j-1)-1} |B_{ik}| + |B_{i(2j-1)}| = \sum_{k=1}^{2j-1} |B_{ik}|.$$

Similarly, the elements in bin $B_{i(2j)}$ have rank at least $\sum_{k=1}^{2j-1} |B_{ik}| + 1$ and at most $\sum_{k=1}^{j} |B_{(i-1)k}| = \sum_{k=1}^{2j} |B_{ik}|$. □

Lemma 2. *$\lfloor \frac{n}{2^i} \rfloor \leq |B_{ij}| \leq \lceil \frac{n}{2^i} \rceil$ for $0 \leq i \leq m$ and $1 \leq j \leq 2^i$*

Proof. We prove by induction that in any round i the sizes of any two bins differ by at most 1, i.e., $\big||B_{ij}| - |B_{ik}|\big| \leq 1$ for $0 \leq i \leq m$ and $1 \leq j, k \leq 2^i$. The statement of the lemma then follows since by an averaging argument and the integrality of the bin sizes, the size of each bin must by of size either $\lceil \frac{n}{2^i} \rceil$ or $\lfloor \frac{n}{2^i} \rfloor$.

For $i = 1$ the n elements of B are partitioned either into two equal sized bins if n is even, or into two bins whose sizes differ by 1 if n is odd.

Now assume that the statement holds for $i - 1$. Take two bins $B_{(i-1)j}$ and $B_{(i-1)k}$. We distinguish two cases.

Case 1. $B_{(i-1)j}$ and $B_{(i-1)k}$ have the same size c. If c is even, then both bins get split up into two bins each and the resulting four bins all have the same size. If c is odd, then each of the bins gets split up into two bins of sizes $\lfloor \frac{c}{2} \rfloor$ and $\lceil \frac{c}{2} \rceil$, respectively, which differ by 1.

Case 2. Without loss of generality, $|B_{(i-1)j}| = c$ and $|B_{(i-1)k}| = c + 1$. If c is even, then $B_{(i-1)j}$ gets split up into two bins both of size $\frac{c}{2}$ and $B_{(i-1)k}$ gets split up into two bins of size $\frac{c}{2}$ and $\frac{c}{2} + 1$, respectively. If c is odd, then $B_{(i-1)j}$ gets split up into two subsets of size $\frac{c+1}{2}$ and $\frac{c+1}{2} - 1$, respectively, and $B_{(i-1)k}$ gets split up into two bins of size $\frac{c+1}{2}$. In any case the bins differ in size by at most 1. □

Lemma 3. *In m rounds the algorithm* ASort *performs less than $6nm$ comparisons.*

Proof. The algorithm by Blum et al. [1] needs at most $5.73n$ comparisons to find the median of n elements and to partition the elements with respect to the median. In the i'th round ASort partitions the elements in every bin $B_{ij}, 1 \leq j \leq 2^i$ with respect to their median. Thus the i'th round needs at most

$$\sum_{j=1}^{2^i} 5.73|B_{ij}| = 5.73 \sum_{j=1}^{2^i} |B_{ij}| = 5.73n \leq 6n$$

comparisons. As the algorithm runs for m rounds the overall number of comparisons is less than $6nm$. □

Theorem 1. *Let $r = \frac{n^2}{\nu(n)}$. Any ranking consistent with the ordering of the bins computed by* ASort *in $\log \nu(n) + 1$ rounds, i.e., with less than $6n(\log \nu(n) + 1)$ comparisons, has a Spearman's footrule distance of at most r to the actual ranking of the elements from B.*

Proof. Using the definition of Spearman's footrule metric and Lemmas 1 and 2 we can conclude that the distance of the ranking of the elements in B to any ranking consistent with the ordering of the bins computed by ASort in m rounds can be bounded by

$$\begin{aligned}
\sum_{j=1}^{2^m} \frac{|B_{mj}|^2}{2} &\leq 2^m \frac{(\lceil \frac{n}{2^m} \rceil)^2}{2} \\
&\leq 2^{m-1} \left(\frac{n}{2^m} + 1 \right)^2 \\
&\leq 2^{m-1} \left(\frac{2n}{2^m} \right)^2, \text{ since } 2^m \leq n \\
&= \frac{n^2}{2^{m-1}}.
\end{aligned}$$

Plugging in $\log \nu(n) + 1$ for m gives a distance less than r as claimed in the statement of the theorem. The claim for the number of comparisons follows from Lemma 3. □

3 Lower Bound

For $r > 0$, by $B_D(\mathrm{id}, r)$ we denote the ball centered at id of radius r with respect to the Spearman's footrule metric, so

$$B_D(\mathrm{id}, r) := \{\pi \in S_n : D(\pi, \mathrm{id}) \leq r\}.$$

Next we estimate the number of permutations in a ball of radius r.

Lemma 4.

$$\left(\frac{r}{en}\right)^n \leq |B_D(\mathrm{id}, r)| \leq \left(\frac{2e(r+n)}{n}\right)^n.$$

Proof. Every permutation $\pi \in S_n$ is uniquely determined by the sequence $\{\pi(i) - i\}_i$. Hence, for any sequence of non-negative integers d_i, $i = 1, \ldots, n$, there are at most 2^n permutations $\pi \in S_n$ satisfying $|\pi(i) - i| = d_i$.

If $d_D(\pi, \mathrm{id}) \leq r$, then $\sum_i |\pi(i) - i| \leq r$. Since the number of sequences of n non-negative integers whose sum is at most r is $\binom{r+n}{n}$, we have

$$|B_D(\mathrm{id}, r)| \leq \binom{r+n}{n} 2^n \leq \left(\frac{2e(r+n)}{n}\right)^n.$$

Next, we give a lower bound on the size of $B_D(\mathrm{id}, r)$. Let $s := \lceil \frac{n^2}{r} \rceil$, and let us first assume that n is divisible by s. We divide the index set $[n]$ into s blocks of size n/s, such that for every $i \in \{1, 2, \ldots, s\}$ the ith block consists of elements $(i-1)\frac{n}{s} + 1, (i-1)\frac{n}{s} + 2, \ldots, i\frac{n}{s}$. For every s permutations $\pi_1, \pi_2, \ldots, \pi_s \in S_{n/s}$ we define the permutation $\rho \in S_n$ to be the concatenation of the permutations applied to corresponding blocks, so $\rho := \pi_1(b_1)\pi_2(b_2) \ldots \pi_s(b_s)$. Note that the distance of ρ to id with respect to Spearman's footrule metric is at most $n \cdot n/s \leq r$, since $|\rho(i) - i| \leq n/s$, for every $i \in [n]$. Obviously, for every choice of $\pi_1, \pi_2, \ldots, \pi_s$ we get a different permutation ρ, which means that we have at least

$$\left(\left(\frac{n}{s}\right)!\right)^s \geq \left(\frac{r}{en}\right)^n$$

different permutations in $B_D(\mathrm{id}, r)$.

If n is not divisible by s, we divide $[n]$ into s blocks of size either $\lceil n/s \rceil$ or $\lfloor n/s \rfloor$, again apply an arbitrary permutation on each of them and we can obtain the same bound in an analogous fashion. □

Theorem 2. *Let $\mathcal{A}$ be a randomized approximate sorting algorithm, let $\nu = \nu(n)$ be a function, and let $r = r(n) = \frac{n^2}{\nu(n)}$.*

If for every input permutation $\pi \in S_n$ the expected Spearman's footrule distance of the output to id *is at most r, then the algorithm performs at least $n\,(\min\{\log \nu, \log n\} - 6)$ comparisons.*

Proof. Let k be the smallest integer such that $\mathcal{A}$ performs at most k comparisons for every input. For a contradiction, let us assume that $k < n\,(\min\{\log \nu, \log n\} - 6)$.

First, we are going to prove

$$\frac{1}{2} n! > 2^k \left(\frac{2e(2r+n)}{n}\right)^n. \tag{1}$$

Since $\log \nu - 6 > k/n$, we have $\frac{\nu}{2^6} > 2^{k/n}$ and since $\nu = \frac{n^2}{r}$ we get

$$\frac{n}{2e} > 2^{k/n} \frac{2e \cdot 2r}{n}. \tag{2}$$

On the other hand, from $\log n - 6 > k/n$ we get $\frac{n}{2^6} > 2^{k/n}$ implying

$$\frac{n}{2e} > 2^{k/n}\frac{2e \cdot n}{n}. \tag{3}$$

Putting (2) and (3) together, we obtain

$$\frac{n}{e} > 2^{k/n}\frac{2e(2r+n)}{n}.$$

Hence

$$\frac{1}{2}n! \geq \left(\frac{n}{e}\right)^n > 2^k \left(\frac{2e(2r+n)}{n}\right)^n,$$

proving (1).

By R we denote the source of random bits for $\mathcal{A}$. One can see R as the set of all infinite 0-1 sequences, and then the algorithm is given a random element of R along with the input. For a permutation $\pi \in S_n$ and $\alpha \in R$, by $\mathcal{A}(\pi, \alpha)$ we denote the output of the algorithm with input π and random bits α.

We fix $\widetilde{\alpha} \in R$ and run the algorithm for every permutation $\pi \in S_n$. Note that with the random bits fixed the algorithm is deterministic. For every comparison made by the algorithm there are two possible outcomes. We partition the set of all permutations S_n into classes such that all permutations in a class have the same outcomes of *all the comparisons* the algorithm makes. Since there is no randomness involved, we have that for every class C there exists a $\sigma \in S_n$ such that for every $\pi \in C$ we have $\mathcal{A}(\pi, \widetilde{\alpha}) = \sigma \circ \pi$. In particular, this implies that the set $\{\mathcal{A}(\pi, \widetilde{\alpha}) : \pi \in C\}$ is of size $|C|$. On the other hand, since the algorithm in this setting is deterministic and the number of comparisons of the algorithm is at most k, there can be at most 2^k classes. Hence, each permutation in S_n is the output for at most 2^k different input permutations. From Lemma 4 we have $|B_D(\mathrm{id}, 2r)| \leq \left(\frac{2e(4r+n)}{n}\right)^n$, and this together with (1) implies that at least

$$n! - 2^k \left(\frac{2e(2r+n)}{n}\right)^n > \frac{1}{2}n!$$

input permutations have output at distance to id more than $2r$.

Now, if both the random bits $\alpha \in R$ and the input permutation $\pi \in S_n$ are chosen at random, the expected distance of the output $\mathcal{A}(\pi, \alpha)$ to id is more than r. Therefore, there exists a permutation π_0 such that for a randomly chosen $\alpha \in R$ the expected distance $d_D(\mathcal{A}(\pi_0, \alpha), \mathrm{id})$ is more than r. Contradiction. □

4 Conclusion

Motivated by an application in market research we studied the problem to approximate a ranking of n items. The metric we use to compare rankings is Spearman's footrule metric, which is within a constant factor to Kendall's tau metric that is frequently used in marketing research. We showed that any randomized algorithm needs at least $n\,(\min\{\log \nu(n), \log n\} - 6)$ comparisons to approximate

a given ranking of n items within expected distance $n^2/\nu(n)$. This result is complemented by an algorithm that shows that $6n(\log \nu(n) + 1)$ comparisons are always sufficient.

In particular, this means that in some cases substantially less comparisons have to be performed than for sorting exactly, provided that a sufficiently large error is allowed. That is, as long as the desired expected error is of order $n^{2-\alpha}$ for constant α one needs $\Omega(n \log n)$ comparisons, which asymptotically is not better than sorting exactly. But to achieve expected error of order $n^{2-o(1)}$ only $o(n \log n)$ comparisons are needed.

Acknowldegments. We are indebted to Jiří Matoušek for comments and insights that made this paper possible.

References

1. M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan. Linear time bounds for median computations. In *STOC '72: Proceedings of the fourth annual ACM symposium on Theory of computing*, pages 119–124. ACM Press, 1972.
2. B. Chazelle. The soft heap: An approximate priority queue with optimal error rate. *Journal of the ACM*, 47(6):1012–1027, 2000.
3. T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. The MIT Press/McGraw-Hill, 1990.
4. P. Diaconis and R. L. Graham. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society*, 39(2):262–268, 1977.
5. H. K. Hwang, B. Y. Yang, and Y. N. Yeh. Presorting algorithms: an average-case point of view. *Theoretical Computer Science*, 242(1-2):29–40, 2000.
6. J. Kahn and J. H. Kim. Entropy and sorting. In *STOC '92: Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 178–187. ACM Press, 1992.
7. D. E. Knuth. *The Art of Computer Programming*, volume 3. Addison Wesley, 1973.
8. C. L. Mallows. Non-null ranking models. *Biometrica*, 44:114–130, 1957.