The List Update Problem – Improved Bounds for the Counter Scheme

Hadas Shachnai
Dept. of Computer Science
The Technion – IIT
Haifa, 32000 ISRAEL
e-mail:hadas@cs.technion.ac.il

Micha Hofri
Dept. of Computer Science
Rice University
Houston TX 77005
e-mail: hofri@cs.rice.edu

Abstract

We consider the problem of dynamic reorganization of a linear list, where requests for the elements are generated randomly with fixed, unknown probabilities. The objective is to obtain the smallest expected cost per access. It has been shown, that when no *a-priori* information is given on the reference probabilities, the *Counter Scheme (CS)* provides an optimal reorganization rule, which applies to *all* possible distributions. In this paper we show that for a list of n elements, arbitrary probabilities and any $\alpha \in (0,1)$, the cost under CS approaches the minimal expected cost up to a ratio of $1+\alpha$ in $O(n \lg n/\alpha^2)$ reorganization steps.

1 Introduction

The list update problem was introduced by McCabe [7]: A fixed set of items is maintained as (an unsorted) linear list or as a serial file. Each request for an item requires a sequential search. The cost of accessing an item is determined by the length of this search. The list may be rearranged during a sequence of requests, so as to achieve a lower average access cost in subsequent requests.
Assuming that each element may be accessed at any time with fixed probability, our goal is to arrange the elements 'correctly,' i.e. in decreasing order of their access probabilities.

A comprehensive survey of many *permutation rules* suggested for the list management and their probabilistic analyses appears in [2]. In this paper we focus on the *Counter Scheme (CS)*, under which the list items are kept in decreasing order by their reference counts, that are updated after every access to the list, i.e., the counters are used as estimates for the unknown access probabilities.

In [5] it was shown that under the above conditions, the CS produces the least expected cost of access *at any time*. Thus, the CS is optimal among all reorganization rules that use no *a-priori* knowledge on the access probabilities. Compared with the optimal static order, the CS was shown in [4] to approach it within a factor of $1+\alpha$, for any $\alpha > 0$, in a number of reorganization steps that is in $O(n^2)$, where n is the number of records in the list.

In the present work we improve this bound and show, that, in fact, the expected access cost under the CS achieves a ratio of $1+\alpha$ to the minimum within $O(n \lg n)$ reorganization steps. This agrees with numerical studies, shown in part in [4].

2 Preliminaries and Notation

We consider is a linear list of n records, $L = \{R_1, \ldots, R_n\}$. Each record R_i is uniquely identified by a key $K_i, 1 \leq i \leq n$. Requests for records are drawn from a multinomial distribution specified by the reference probability vector (*rpv*): $\bar{p} = (p_1, \ldots, p_n)$. Thus, R_i may be requested at each access with probability p_i. This is known as the *independent reference model (irm)* [4, 8]. We assume w.l.o.g. a renumbering of the records, such that $p_1 \geq \cdots \geq p_n$.

Each reference requires a sequential search of the list. We define C, the cost of a single access, as the number of key-comparisons made till the specified record is reached. Under the *irm*, with a fixed *rpv*, the average access cost to the list is minimized when the records are in the optimal static order: R_i precedes R_j whenever $p_i > p_j$. Getting there requires a complete knowledge of the *rpv*, or at least of the relative magnitude of the access probabilities. This knowledge is assumed unavailable.

The initial arrangement of L is assumed to be randomly selected (with equal probability) out of all its possible permutations. As the list is referenced, it is constantly reorganized, with the aim of approaching the optimal ordering as the reference sequence grows longer.

In this work we derive results for the CS. Our performance measure is the expected access cost after the mth reference, $m \geq 0$, denoted by $C_m(CS \mid \bar{p})$.

Let σ_m denote the order of the list elements after the mth reference: $\sigma_m(i)$ is the position of R_i in the list. Let $\text{Prob}_{CS}(\sigma_m(j) < \sigma_m(i))$ be the probability that R_j precedes R_i after the mth reference, when the list is reorganized by the CS. Then we can write the expected access cost under the CS after the mth reference as

$$C_m(CS|\bar{p}) = C(OPT \mid \bar{p}) + \sum_{1 \leq i < j \leq n} (p_i - p_j) \cdot \text{Prob}_{CS}(\sigma_m(j) < \sigma_m(i)), \tag{1}$$

where

$$C(OPT \mid \bar{p}) = \sum_{i=1}^{n} i p_i = 1 + \sum_{1 \leq i < j \leq n} p_j, \tag{2}$$

is the expected access cost under the optimal static arrangement of the list. All the index pairs $(i < j)$ appear exactly once in this summation.

By the strong Law of Large Numbers [6], for any $p_i > p_j$,

$$\lim_{m\to\infty} \mathrm{Prob}_{CS}(\sigma_m(j) < \sigma_m(i)) = 0.$$

Hence

$$\lim_{m\to\infty} C_m(CS|\bar{p}) = C(OPT \mid \bar{p}).$$

What the LLN does not provide is an estimate of the rate of convergence of $C_m(CS|\bar{p})$ to its limit. We would like to compute m, the number of references (= reorganization steps, in the CS scheme) so that $C_m(CS|\bar{p})$ is close enough to $C(OPT \mid \bar{p})$, for any $\bar{p}$.

The following lemma formalizes the notion of a stopping point for the reorganization process under the CS.

Lemma 1: [4] The cost function $C_m(CS|\bar{p})$ is monotone decreasing in m for all $m \geq 1$.

Hence, given some $\alpha > 0$, once we find a number of steps, m^*, such that

$$C_{m^*}(CS|\bar{p}) \leq (1+\alpha)C(OPT \mid \bar{p}). \tag{3}$$

then for all $m > m^*$, also $C_m(CS|\bar{p}) \leq (1+\alpha)C(OPT \mid \bar{p})$.

The following lemma gives the desired stopping point when the access distribution is *assumed known* (at least up to the mapping of probabilities to the keys). In the next section we use this result to derive a *distribution free* bound on the stopping point. The lemma is based on the bound shown in [1]:

$$\mathrm{Prob}_{CS}(\sigma_m(j) < \sigma_m(i)) \leq (1-(\sqrt{p_i}-\sqrt{p_j})^2)^m.$$

Lemma 2: [1] For a given *rpv* $\bar{p}$ and any $0 < \alpha < 1$, the cost under CS achieves a ratio of $(1+\alpha)$ to $C(OPT \mid \bar{p})$ within m^* steps, where

$$m^* = \min_{m\geq 1}\{m \mid \sum_{1\leq i<j\leq n} (p_i - p_j)(1-(\sqrt{p_i}-\sqrt{p_j})^2)^m \leq \alpha(1+\sum_{1\leq i<j\leq n} p_j)\}. \tag{4}$$

3 A Stopping Point for the CS

The following lemma is the crux of the present result:

Lemma 3: For $0 < \alpha < 1$ and *rpv* $\bar{p}$, such that $C(OPT \mid \bar{p}) \geq 1 + 1/n^r$, $C_m(CS|\bar{p})$ approaches the optimal cost to within a factor $1+\alpha$ following $O(\alpha^{-2} n \lg n)$ reorganization steps.

Note: The value r is a parameter of the proof, as we see below. While such an r exists for *any* reference distribution, it is noteworthy that for most $\bar{p}$ the optimal cost is substantially larger than 2.

Proof: The condition $C(OPT \mid \bar{p}) \geq 1 + 1/n^r$ implies the inequality

$$\sum_{1 \leq i < j \leq n} p_j \geq n^{-r}. \tag{5}$$

Define the set of ordered pairs

$$S = \{(i,j) \mid 1 \leq i < j \leq n, \quad p_i - p_j > \alpha p_j\}. \tag{6}$$

By equation (4) it suffices to find the minimal $m \geq 1$, such that[1]

$$\sum_{1 \leq i < j \leq n} (p_i - p_j)(1 - (\sqrt{p_i} - \sqrt{p_j})^2)^m \leq \alpha(1 + \sum_{(\dot{i},\dot{j}) \in S} p_j + \sum_{(\dot{i},\dot{j}) \notin S} p_j). \tag{7}$$

Instead, we proceed to find the value of $m \geq 1$ satisfying

$$\sum_{1 \leq i < j \leq n} (p_i - p_j)(1 - (\sqrt{p_i} - \sqrt{p_j})^2)^m \leq \alpha \left(1 + \sum_{(\dot{i},\dot{j}) \notin S} p_j + \max(\sum_{(\dot{i},\dot{j}) \in S} p_j, n^{-r}) \right), \tag{8}$$

and then convert it to our needs. From relation (5) it follows, that whether $\sum_{(\dot{i},\dot{j}) \in S} p_j < n^{-r}$ or the reverse holds, it is true that

$$\begin{aligned} 1 + \sum_{(\dot{i},\dot{j}) \notin S} p_j + \max(\sum_{(\dot{i},\dot{j}) \in S} p_j, n^{-r}) &\leq 1 + \sum_{(\dot{i},\dot{j}) \in S} p_j + 2 \sum_{(\dot{i},\dot{j}) \notin S} p_j \\ &\leq 1 + 2 \sum_{1 \leq i < j \leq n} p_j. \end{aligned}$$

Hence, once we have found an m that satisfies relation (8), that m also satisfies

$$\sum_{1 \leq i < j \leq n} (p_i - p_j)(1 - (\sqrt{p_i} - \sqrt{p_j})^2)^m \leq \alpha(1 + 2 \sum_{1 \leq i < j \leq n} p_j),$$

and *a-fortiori* it satisfies

$$C_m(CS|\bar{p}) \leq C(OPT \mid \bar{p})(1 + 2\alpha). \tag{9}$$

[1] We use dots under indices that take part in the summation, when it may not be obvious.

To obtain m^* from relation (8) we use the definition (6), dropping on both sides the contributions of pairs not in S (which satisfy the inequality for any $m \geq 0$) – and also an extra α on the right-hand side, and are left with the requirement on m:

$$\sum_{(i,j)\in S} (p_i - p_j)(1-(\sqrt{p_i}-\sqrt{p_j})^2)^m \leq \alpha \max(\sum_{(i,j)\in S} p_j,\ n^{-r}). \tag{10}$$

Since $\max(\sum_{(i,j)\in S} p_j,\ n^{-r})$ is at least n^{-r}, we tighten the requirement on m by using from now on the relation

$$\sum_{(i,j)\in S} (p_i - p_j)(1-(\sqrt{p_i}-\sqrt{p_j})^2)^m \leq \alpha n^{-r}. \tag{11}$$

The following notation is useful:

$$V = \sum_{(i,j)\in S} p_i, \tag{12}$$

$$N \equiv |S|, \quad \text{and} \quad A \equiv \sum_{(i,j)\in S} (p_i - p_j). \tag{13}$$

We note that

$$\sum_{(i,j)\in S} p_j = V - A. \tag{14}$$

Claim 1: For any $(i,j) \in S$

$$(\sqrt{p_i}-\sqrt{p_j})^2 > \left(1 - \frac{2}{1+\sqrt{1+\alpha}}\right)(p_i - p_j). \tag{15}$$

Proof: Let $p_i = q^2 p_j$. For $(i,j) \in S$, definition (6) requires that $q^2 > 1+\alpha$. Compute

$$\frac{(\sqrt{p_i}-\sqrt{p_j})^2}{p_i - p_j} = \frac{p_i + p_j - 2\sqrt{p_i p_j}}{p_i - p_j} = 1 - \frac{2}{q+1},$$

which yields inequality (15). □

Let $d \equiv 1 - \frac{2}{1+\sqrt{1+\alpha}}$. Using relation (15), the left-hand side of relation (11) is bounded by $\sum_{(i,j)\in S}(p_i - p_j)(1-d(p_i - p_j))^m$. We simplify the task of finding an upper bound for m, by "maximizing" this last expression. Specifically, while $A = \sum_{(i,j)\in S}(p_i - p_j)$, we consider

all sets $\{a_{ij} \geq 0\}$ such that $\sum_{(i,j)\in S} a_{ij} = A$, and look for one that with any given $m > 1$, maximizes the function

$$\sum_{(i,j)\in S} a_{ij}(1 - da_{ij})^m.$$

The maximum is obtained when $a_{ij} = \frac{A}{N}$ for all $(i,j) \in S$. Thus, it is sufficient to find the minimal m satisfying

$$\sum_{(i,j)\in S} \frac{A}{N}\left(1 - \frac{dA}{N}\right)^m < \alpha n^{-r}. \tag{16}$$

Getting rid of the sum, it remains to resolve the minimal m such that

$$\left(1 - \frac{dA}{N}\right)^m < \frac{\alpha}{n^r A}. \tag{17}$$

We can write now an expression for m^*, but we need first to relate the values of A and N to the problem parameters, n and α.

Claim 2: For any $n \geq 2$ and A,N as defined in (13), $A/N \geq \alpha/n(1+\alpha)$.

Proof: Let

$$N_k \equiv \left|\{(i,j) \in S \,:\, j = k\}\right| \quad \text{and} \quad V_k \equiv \sum_{(i,k)\in S} p_i.$$

That is, for $2 \leq k \leq n$, N_k is the size of the subset of ordered pairs in S, in which the smaller probability is p_k. Clearly

$$N = \sum_{k=1}^{n} N_k \quad \text{and} \quad V = \sum_{k=1}^{n} V_k.$$

In addition, for any $(i,j) \in S$, the condition $p_i - p_j > \alpha p_j$ leads to

$$p_i - p_j > \alpha p_i/(1+\alpha). \tag{18}$$

Hence,

$$\frac{A}{N} = \frac{\sum_{k=1}^{n}\sum_{(i,k)\in S}(p_i - p_k)}{\sum_{k=1}^{n} N_k} > \min_{1\leq k\leq n} \frac{\sum_{(i,k)\in S} p_i\alpha/(1+\alpha)}{N_k} = \frac{\alpha}{1+\alpha}\min_{1\leq k\leq n}\frac{V_k}{N_k}.$$

The order $p_1 \geq \cdots \geq p_n$ implies that if $(i,k) \in S$, then for all $1 \leq i' < i$ also $(i',k) \in S$. Therefore

$$\frac{V_k}{N_k} = \frac{p_1 + \cdots + p_{N_k}}{N_k},$$

i.e., the ratio V_k/N_k is the average of the N_k largest probabilities. This value is at least $1/n$. □

The last relation we need:

Claim 3: For any n and $\bar{p}$, $A < n$.

Proof: Clearly $A = \sum_{(i,j)\in S}(p_i - p_j) \le A' \equiv \sum_{1\le i<j\le n}(p_i - p_j) = \sum_{i=1}^{n}(n-2i+1)p_i$ which is at most $n-1$. □

Let $s \equiv \frac{\alpha}{1+\alpha}$. Using (17) and the above claims we can bound m^* by solving the inequality

$$\left(1 - \frac{s \cdot d}{n}\right)^m < \frac{\alpha}{n^{r+1}}, \tag{19}$$

and $m = \frac{n}{sd}\ln(\frac{n^{r+1}}{\alpha})$ provides the desired result, with $sd = 1 + (1-2\sqrt{1+\alpha})/(1+\alpha)$.

Referring back to equation (9), we need to replace α by $\alpha/2$. Setting

$$m^* = \frac{n(r+1)(2+\alpha)(4+\alpha+2\sqrt{4+2\alpha})}{\alpha^2}\ln(\frac{2n}{\alpha}) \tag{20}$$

produces the statement of the lemma. □

Comment: The role of the parameter r above deserves a discussion. In one sense, it is a mere technical device: for any value of $C(OPT \mid \bar{p})$ there is an r such that $C(OPT \mid \bar{p}) \ge 1 + n^{-r}$, and the above proof holds. In fact, for any *rpv* likely to arise in practice, $r = 0$ satisfies the condition. However, it is easy to manufacture a sequence of *rpv*s such that $C(OPT \mid \bar{p})$ gets arbitrarily close to 1, and "requires" larger and larger values of r. The simplest example is $p_2 = 1 - p_1 = n^{-r}$, and all other $p_j = 0$ (to avoid trivialities with records that are never requested, we may assume these p_j are all equal to n^{-2r}). This sequence suggests that our bound is not really "distribution-free" as we would like it to be. The only cases where this occurs concern such skewed distributions that the entire issue of reorganizing the list to improve its access time is nearly meaningless. Hence, while this is a real feature of the reorganization problem, it appears to have no practical significance.

Based on the last lemma and comment, we can state

Theorem 1: For any $0 < \alpha < 1$ and *rpv* $\bar{p}$, $C_m(CS|\bar{p})$ approaches the optimal cost to within a factor $1+\alpha$ following finite, precomputable number of reorganization steps, $m^*(\alpha, \bar{p})$. m^* is proportional to $n\lg n$, to α^{-2}, and for vectors $\bar{p}$ that are nearly concentrated in a single record, also to $\log(C(OPT \mid \bar{p}) - 1)^{-1}/\log n$. □

4 Concluding Remarks

The theorem above resolves a discrepancy that was evident in [4], between the bounds we could prove and the numerical evidence. In fact, experiments for the Zipf function and geometric distributions showed that the required number of references to organize the list followed the $n \log n$ pattern. Our satisfaction at resolving the issue is marred by the surprising fact that there appears to be no truly universal, distribution-free bound. Note that the example provided in the comment above, of the *rpv*s with $C(OPT \mid \bar{p})$ that approach 1, indicate that the issue is not with our proof method, but that the difficulty is inherent in the reorganization process under the counter scheme (or any other method that only moves a record once it is referenced; when $p_2 = n^{-r}$, the record R_2 is only requested once in every n^r accesses, on the average!).

On the other hand, the difficulty arises in pathological cases only, and for all practical purposes the expression we proposed for m^* exhibits the behavior of the needed number of references to achieve the desired goal. We use the term 'behavior' rather than 'value.' We may expect that the given m^* much exceeds the bound required for all but isolated types of distributions. In [4] we argue why $C_m(CS|\bar{p})$ usually converges quite promptly to values close to $C(OPT \mid \bar{p})$, and give some numerical examples.

By way of apologizing for our notation we should mention that although we use the Big-O notation borrowed from asymptotics, the linear list scheme is only meaningful for short to moderate lists.

References

[1] Cohen A., Rabinovich Y., Schuster A., Shachnai H., "Optimal Bounds on Tail Probabilities - A Simplified Approach", TR #911, Technion IIT, CS Dept., May 1997.

[2] Hester J.H., Hirschberg D.S., "Self-Organizing Linear Search", *ACM Comput. Surv.*, 17, 3, pp. 295-312, 1985.

[3] Hofri M., *Analysis of Algorithms*, Oxford University Press, 1995.

[4] Hofri M., Shachnai H., "Self-Organizing Lists and Independent References - a Statistical Synergy", *Jour. of Alg.*, 12, 533-555, 1991.

[5] Hofri M., Shachnai H., "On the Optimality of Counter Scheme for Dynamic Linear Lists", *Inf. Process. Lett.*, 37, pp. 175-179, 1991.

[6] Karr A.F., *Probability*, Springer-Verlag, 1992.

[7] McCabe J., "On Serial Files with Relocatable Records", *Operations Research*, 13, pp. 609-618, 1965.

[8] Topkis D. M., "Reordering Heuristics for Routing in Communication Network", *J. of Applied Prob.*, pp. 130-143, 1986.