Rates of convergence for a self-organizing scheme for binary search trees

[short title: Convergence rates for self-organizing trees]

BY ROBERT P. DOBROW AND JAMES ALLEN FILL[1]

The Johns Hopkins University

Abstract

The move-to-root heuristic is a self-organizing rule which attempts to keep a binary search tree in near-optimal form. It is a tree analogue of the move-to-front scheme (also known as the weighted random-to-top card shuffle or Tsetlin library) for self-organizing lists. We study convergence of the move-to-root Markov chain to its stationary distribution and show that move-to-root converges two to four times faster than move-to-front for many examples. We also discuss asymptotics for expected search cost. For equal weights, $cn/\ln n$ steps are necessary and sufficient to drive maximum relative error to 0.

[1]Research for both authors supported by NSF grant DMS-9311367.

[2]*AMS 1991 subject classifications.* Primary 60J10; secondary 68P10, 68P05.

[3]*Keywords and phrases.* Markov chains, convergence to stationarity, self-organizing search, binary search trees, move-to-root rule, move-to-front rule.

1 Introduction and Summary

Suppose that a collection of n records is arranged in a sequential list. Associated with the ith record is a weight y_i measuring the frequency of its use. We assume that $y_i > 0$ and normalize so that $\sum y_i = 1$. At each unit of time, independently of all other moves, item i is removed from the list with probability y_i and replaced at the front of the list. This process, which gives a Markov chain on the permutation group S_n, is known as the move-to-front (MTF) heuristic for self-organizing lists. Other names are the weighted random-to-top card shuffle, the Tsetlin library, and the heaps process. Background references for this model include Rivest (1976), Bitner (1979), Hendricks (1989), and Fill (1993). Fill (1993) and Diaconis (1993) study rates of convergence for MTF under various assumptions on the weights.

There has been recent interest in self-organizing rules for other data structures. In particular, the binary search tree is a commonly used structure which exploits the ordering of records to achieve faster search time. Records are stored at the nodes of a tree in such a way that a traversal of the tree produces the records in their linear order.

Allen and Munro (1978) introduced the move-to-root (MTR) heuristic for binary search trees. Dobrow and Fill (1993) show that the Markov chain for MTR can be derived by lumping the Markov chain for MTF and determine numerous characteristics of the chain.

In Section 2 we review binary search trees and give some of the relevant results from previous work. (See Dobrow and Fill (1993) for more details.) In Section 3 we analyze the convergence of the MTR chain to its stationary distribution. Roughly, the results are as follows. For most weight classes, if k steps are necessary and sufficient for convergence of MTF with respect to total variation distance, then $k/2$ steps are sufficient for MTR and $k/4$ steps are necessary. These bounds are gotten by common techniques: Sufficiency is proven by exhibiting a natural coupling of the MTR chain. Necessity is shown by identifying a specific event from which variation distance can be suitably bounded from below. We have been unable to bridge the gap between $k/4$ and $k/2$.

In Section 4 we study convergence of expected search cost, the average cost of accessing a record. In the case of equal weights ($p_i \equiv 1/n$) we show that $cn/\ln n$ steps are necessary and sufficient to drive the maximum relative error to 0.

2 Binary search trees and move-to-root

A *binary tree* is a finite tree with at most two "children" for each node and in which each child is distinguished as either a left or right child. Consider a binary tree in which the nodes are labeled with elements of some linearly ordered set. Inorder traversal is a common method for traversing the tree: visit the root after visiting the left subtree and before visiting the right subtree. If this traversal yields the labels in order, the tree is called a *binary search tree.* For example, the set of all binary search trees on 3 nodes is given by:

Figure 1.

Consider the ordered, indexed set of $[n] := \{1, 2, \ldots, n\}$ of n records. Let B_n be the set of all labeled binary search trees on n nodes. It is easy to show that $|B_n| = \binom{2n}{n} /(n+1)$, which is the nth Catalan number. In what follows we use the term "tree" for binary search tree.

The *move-to-root* (MTR) operation is defined as a series of simple exchanges between nodes. A *simple exchange* for a requested record j is as follows:

(i) Do nothing if j is the root.

(ii) If j is the left child of its parent m, the resulting tree will be the same as the original except for the subtree whose root was m. Record j is "rotated" up to m so that j becomes the root of this subtree. The old left subtree of j doesn't change in relation to j. The old right subtree of j becomes the left subtree of m. The old right subtree of m keeps its relation to m. The transformation is best understood by examining Figure 2-L.

(iii) If j is the right child of m, perform the analogous transformation. (See Figure 2-R.)

The MTR operation performs a sequence of simple exchanges until the requested record is moved to the root of the tree. For n-node trees it is easily shown that the sequence of operations generated by MTR gives an ergodic

(aperiodic, irreducible, and positive recurrent) Markov chain on the space B_n.

Figure 2.

For $T \in B_n$ and $i \neq j$, we say that i is an *ancestor* of j in T, and write $i <_a^T j$, if j is an element of the subtree which has i as its root. A tree is uniquely determined by its ancestry relations.

A key observation for analyzing MTR is based on Lemma 3.2 in Allen and Munro (1978), which we reproduce, with a slight extension:

Lemma 2.1 *Suppose record i has been requested at least once in a tree modified according to MTR. If $i < j$, then $i <_a j$ if and only if the most recent request for i has occurred since the most recent request (if any) for any of $i+1, \ldots, j$. Similarly, if $i > j$, then $i <_a j$ if and only if the most recent request for i has occurred since the most recent request for any of $j, \ldots, i-1$.*

Proof When either simple exchange or MTR is used, if i is requested then i is the only record which *becomes* an ancestor of any records. Also, i will *cease* to be an ancestor of $j > i$ if and only if an element k, where $i < k \leq j$, is requested. This gives the first part of the lemma. The second part is shown similarly. ■

Consider the operation of inserting records into an initially empty tree. This defines a mapping $t : S_n \to B_n$ where for $\sigma \in S_n$, $t(\sigma)$ is the tree obtained by successively inserting records $\sigma_1, \ldots, \sigma_n$ into an empty tree. Dobrow and Fill (1993) show that the MTR chain can be obtained by lumping the MTF chain with respect to the mapping t. In particular,

Theorem 0 *Let Q denote the transition matrix for MTR and P the transition matrix for MTF. Let $\Pi(T)$ denote the set of permutations that are mapped to a given tree T by t. Then for $S, T \in B_n$ and $k \geq 0$,*

$$Q^k(S,T) = \sum_{\sigma \in \Pi(T)} P^k(\pi, \sigma) \text{ for all } \pi \in \Pi(S). \tag{1}$$

3 Convergence rates for MTR

Let $0 < p_1 \leq p_2 \leq \cdots \leq p_n$ be a fixed ordered list of weights with $\sum p_i = 1$. Let $\sigma \in S_n$ and suppose that $\mathbf{y} = \mathbf{p} \circ \sigma$ is such that $y_i = p_{\sigma(i)}$ is the probability of requesting record i at any step of MTR. Let Q_T^k denote the distribution of MTR at time k when the chain is started at tree T. Let Q^∞ denote the stationary distribution of MTR. Note that for fixed $\mathbf{p}$ the quantities Q_T^k and Q^∞ depend on the permutation σ.

Our measure for the distance between the MTR chain at time k and its stationary distribution will be the usual *total variation distance.* We treat the worst initial T:

$$d(k; \mathbf{y}) := \max_{T \in B_n} \|Q_T^k - Q^\infty\|_{TV} = \max_{T \in B_n} \max_{A \subseteq B_n} |Q_T^k(A) - Q^\infty(A)|.$$

For general background on variation distance see Aldous and Diaconis (1986, 1987). Taking maxima over all orderings of the weights leads to the maximum variation distance

$$d(k) := \max_{\sigma \in S_n} d(k, \mathbf{p} \circ \sigma).$$

Given a triangular array of weights $\mathbf{p}_n = (p_{n,i},\ i = 1, \ldots, n),\ n \geq 1$, say that $k = k(n, c)$ steps are *sufficient* for total variation convergence to stationarity if there exists positive constants α and β such that for each fixed c we have $d(k) \leq \alpha e^{-\beta c} + o(1)$ as $n \to \infty$. Say that $k = k(n, c)$ steps are *necessary* for convergence to stationarity if there exists a function h, independent of n, such that $d(k) \geq h(c)$ and $h(c) > 0$ is bounded away from 0. In the case of uniform, Zipf's law, and generalized Zipf's law weights, as discussed below, we will even be able to take $h(c) \to 1$ as $c \to -\infty$. If $k(n, c) = f(n) + cg(n)$ steps are necessary and sufficient for convergence to stationarity where $g(n) = o(f(n))$, we say that a "cutoff" occurs.

We stress that, by our definiton, if k steps are sufficient for MTR under a class of weights, then k steps are sufficient for all orderings of the weights and all initial trees.

For $s > 1$, let $\zeta(s) := \sum_{i=1}^\infty i^{-s}$. Let $\ln^{(i)}$ denote the ith iterated logarithm of n. We will consider the following choices of weights, where each weight class p_i is listed up to the constant of proportionality.

Weights	$p_i \propto$
Uniform	1
Zipf's law	$1/(n-i+1)$
Generalized Zipf's law (GZL)	$1/(n-i+1)^s, \quad s > 0$ fixed
Power law	$i^s, \quad s > 0$
Geometric	$\theta^{n-i}, \quad 0 < \theta < 1$ fixed

The weights we have chosen are standard examples and cover a very wide class. See Knuth (1973) for an interesting discussion of the motivation for using Zipf's and generalized Zipf's law weights. For large n, generalized Zipf's law with $s = \ln 4 / \ln 5 \doteq 0.86$ approximately fulfills the "80–20" rule of thumb that has often been observed for commercial computing applications; this rule states that 80% of the transactions deal with the most active 20% of the file. Generalized Zipf's law weights with s slightly larger than 1 are suggested by Schwartz (1963) as a model for word frequencies. Diaconis (1993) treats uniform, geometric, and generalized Zipf's law weights for MTF. Knuth (1973) discusses the "wedge-shaped" distribution obtained by letting $p_i \propto i$, which we have generalized and dubbed the power law.

Theorem 1 *Table 1 give rates of convergence to stationarity for MTR.*

Table 1.

Weights	Sufficiency	Necessity
Uniform	$\frac{1}{2}n(\ln n + c)$	$\frac{1}{4}n(\ln n - c)$
Zipf's law	$\frac{1}{2}n \ln n(\ln n - \ln^{(2)} n + c)$	$\frac{1}{4}n \ln n(\ln n - \ln^{(2)} n - c)$
GZL		
$0 < s < 1$	$\frac{1}{2}\frac{n}{1-s}(\ln n - \ln^{(2)} n + c)$	$\frac{1}{4}\frac{n}{1-s}(\ln n - \ln^{(2)} n - c)$
$s > 1$	$\frac{\zeta(s)}{2}n^s(\ln n - \ln^{(2)} n + c)$	$\frac{\zeta(s)}{4}n^s(\ln n - \ln^{(2)} n - c)$
Power law	cn^{s+1}	cn^{s+1}
Geometric	$c\theta^{-n}$	$c\theta^{-n}$

Remarks:

1. Since the MTR chain can be derived by lumping the MTF chain, we know that the total variation distance at time k for MTR can never be larger than for MTF. However, our results for uniform, Zipf's law, and GZL weights explicitly quantify the speedup in the rates of convergence: MTR is two to four times as fast as MTF. In the case of geometric and power law weights, no cutoff occurs. We have not investigated speedup in this case; any constant factor can be absorbed in c.

2. Another, more locally sensitive, measure of discrepancy between distributions is *separation*, defined here for initial tree S by

$$s_S(k) := \max_T \left(1 - \frac{Q^k(S,T)}{Q^\infty(T)}\right).$$

For a detailed treatment of separation, which bounds variation distance, see, e.g., Diaconis and Fill (1990).

It is not difficult to show that the worst case (over initial trees and orderings of the weights) separation for MTR is, for any set of weights, equal to the worst case (over initial lists) separation for MTF. This follows from lumpability and Theorem 4.1 in Fill (1993) along with the observation that $\Pi(T)$ is a singleton when T is the tree corresponding to either of the permutations $\mathrm{id} = (1, \ldots, n)$ or $\mathrm{rev} = (n, \ldots, 1)$. Thus, for example, $n \ln n + cn$ steps are necessary and sufficient for separation for MTR in the case of uniform weights.

For MTF, the lead order terms for the number of steps that are necessary and sufficient for convergence are the same for total variation distance and separation. Interestingly, this is not the case for MTR: the discrepancy is at least a factor of 2 for uniform, Zipf's law, and GZL weights.

3.1 Proof of Theorem 1: Sufficiency

Coupling is a probabilistic technique which is useful for bounding variation distance. Let $X = (X_n)$ and $Y = (Y_n)$ be two copies of the MTR chain such that the X-chain has an arbitrary initial distribution and the Y-chain is started in stationarity. A coupling time T is a stopping time such that $X_n = Y_n$ for all $n \geq T$. Total variation distance is bounded above by the tail probability of a coupling time. That is,

$$d(k; \mathbf{y}) \leq P(T > k), \quad k \geq 0. \tag{2}$$

Before describing the coupling we introduce some terminology following Dobrow and Fill (1993). For $R \subseteq [n]$, write $r_1 < r_2 < \cdots < r_m$ for the elements of R. Define $r_0 := 0$ and $r_{m+1} := n + 1$. Let

$$g_i(R) := r_{i+1} - r_i - 1, \quad i = 0, \ldots, m,$$

denote the number of integers in the interval (r_i, r_{i+1}). Then $g_i(R)$ is called the *i-th gap* of R.

Theorem 2 *For an n-node tree suppose record i is requested with probability y_i. Under MTR, let T be the first time that the sequence of selected records has no gap of size greater than 1. Then*

$$d(k; \mathbf{y}) \leq P(T > k) \leq \sum_{i=1}^{n-1} (1 - y_i - y_{i+1})^k. \tag{3}$$

Before proving Theorem 2 we state a lemma. For $T \in B_n$, write T^k for the tree obtained after k steps of MTR.

Lemma 3.1 *Let $i \in [n]$ and let $r_1, r_2, \ldots$ be a sequence of record requests. Suppose $k = \min\{m : r_m = i\}$. Then for all $S, T \in B_n$, $j \in [n]$, and $k' \geq k$,*

(a) $i <_a^{T^{k'}} j$ if and only if $i <_a^{S^{k'}} j$, and

(b) $j <_a^{T^{k'}} i$ if and only if $j <_a^{S^{k'}} i$.

Proof Without loss of generality we can assume that record i is stored at the root node in S and T and that $k = 1$. The lemma is then an easy consequence of Lemma 2.1. ■

We now prove Theorem 2.

Proof Consider the following coupling of the MTR chain: Begin with $X, Y \in B_n$. Now select records according to MTR. When a record is selected, move that record to the root in both X and Y. We claim that at the first time T that the set of records requested at least once has no gap of size greater than 1 and thereafter, the two processes have the same value. Thus, by definition, the coupling time T has the property that if record i has not been requested by time T then records $i - 1$ and $i + 1$ have.

By Lemma 2.1 in Dobrow and Fill (1993), to show that the two trees agree at time T it suffices to show

$$i <_a^{X^T} j \Rightarrow i <_a^{Y^T} j \text{ for all } i, j \in [n].$$

We suppose $i < j$; the case $i > j$ is handled similarly. Suppose $i <_a^{X^T} j$. Let $R = \{r_1, \ldots, r_T\}$ be the set of records requested through time T.

If $i \in R$ then $i <_a^{Y^T} j$ by Lemma 3.1(a). If $i \notin R$ then $i + 1 \in R$ since R has no gaps of size greater than 1. If $j \in R$ then $i <_a^{Y^T} j$ by Lemma 3.1. If $j \notin R$ we will derive a contradiction. Since $j \notin R$ and $i + 1 \in R$, we have $j \neq i + 1$. Thus there exists $i < m < j$ such that $m \in R$. Suppose that m is first requested at time $k' \leq T$. Then $i \not<_a^{X^{k'}} j$. Since $i \notin R$ it follows that $i \not<_a^{X^{k''}} j$ for all integers $k'' \in [k', T]$. In particular, $i \not<_a^{X^T} j$.

The first inequality in (3) follows since T is a coupling time. Let A_i be the event that neither of the records i and $i + 1$ has been requested by time k. Then

$$P(T > k) = P(\cup_{i=1}^{n-1} A_i), \tag{4}$$

and the second inequality in (3) follows from subadditivity.

■

Remarks:

1. In the case of MTF, the first time T' that all but one of the records has been requested is a coupling time. Note that the distribution of our T, unlike that of T', depends on the order σ of the weights.

2. Our coupling time is not a fastest coupling time, for which the inequality in (2) would be an equality. For example, stopping when both trees agree is faster (but still not fastest).

Theorem 3 *For weights $0 < p_1 \leq \cdots \leq p_n$, suppose there exists positive constants α and β such that*

$$\sum_{i=1}^{n-1} (1 - p_i)^{2k} \leq \alpha e^{-\beta c}. \tag{5}$$

Then k steps are sufficient for MTR.

Proof Recalling the notation set forth at the beginning of Section 3, in particular $\mathbf{y} = \mathbf{p} \circ \sigma$, the proof is a direct consequence of Theorem 2 via the

following chain of inequalities:

$$\begin{aligned} d(k;\mathbf{y}) &\leq \sum_{i=1}^{n-1}(1-y_i-y_{i+1})^k \\ &\leq \sum_{i=1}^{n-1}[(1-y_i)(1-y_{i+1})]^k \\ &\leq \left(\sum_{i=1}^{n-1}(1-y_i)^{2k}\right)^{1/2}\left(\sum_{i=1}^{n-1}(1-y_{i+1})^{2k}\right)^{1/2} \\ &\leq \sum_{i=1}^{n-1}(1-p_i)^{2k}. \end{aligned}$$

The third inequality here is from Cauchy–Schwarz. ■

For a fixed class of weights, let k be the number in the sufficiency column of the table in Theorem 1 corresponding to that weight class. To prove sufficiency in Theorem 1 it is enough to show that

$$\sum_{i=1}^{n-1}(1-p_i)^{2k} \leq \alpha e^{-\beta c},$$

for positive constants α and β. For uniform, Zipf's law, and GZL weights, Diaconis (1993) shows that

$$\sum_{i=2}^{n}(1-p_i)^{2k} \leq \alpha e^{-\beta c} \tag{6}$$

for positive constants α and β. It is easy to see for these weights that if k steps are sufficient, then $kp_1 \to \infty$. Thus

$$\begin{aligned} \sum_{i=1}^{n-1}(1-p_i)^{2k} &\leq \sum_{i=2}^{n}(1-p_i)^{2k}+(1-p_1)^{2k} \\ &\leq \sum_{i=2}^{n}(1-p_i)^{2k}+e^{-2kp_1} \\ &= \sum_{i=2}^{n}(1-p_i)^{2k}+o(1) \\ &\leq \alpha e^{-\beta c}+o(1). \end{aligned}$$

A similar analysis can be done for geometric and power law weights. Diaconis treats the former. In the latter case it follows from Diaconis's Theorem 3 that cn^{s+1} steps suffice. In both cases kp_1 is approximately a constant times c, which doesn't alter the results.

Remarks:

1. Diaconis (1993) gives general conditions on a triangular array of weights in order for the cutoff phenomenon to occur for MTF. For such weights the inequality (6) is satisfied and $kp_1 \to \infty$. At this level of generality, if k steps are sufficient for MTF, then $k/2$ steps are sufficient for MTR.

2. Given the factor of 2 discrepancy in Theorem 1, it is natural to investigate the source of error in the two inequalities in (3). We show that in the case of uniform weights the second inequality, at least, is quite sharp. For uniform weights it is not difficult to determine the exact distribution of T of Theorem 2. By considering a variant of the coupon collector's problem it follows that

$$P(T > k) = \sum_{u=1}^{n} P_k(u)P(E_u), \tag{7}$$

where $P_k(u)$ is the probability that throwing k balls into n urns leaves exactly u urns empty and E_u is the event that picking u numbers at random from $[n]$ leaves a gap of size greater than 1. The probability $P(E_u)$ is easily computed, while $P_k(u)$ is well-known from ordinary coupon collecting. We have made a careful asymptotic analysis of (7) and find that the subadditive bound on (4) gives the correct lead order term.

3.2 Proof of Theorem 1: Necessity

From the definition of total variation distance it follows that for any event B and any starting tree T, $d(k)$ is bounded from below by $|Q_T^k(B) - Q^\infty(B)|$.

Let $T = t(\text{rev})$, where rev is the reversal permutation. That is, T has record n at the root, and record $i+1$ is the parent of record i for $i = n-1, \ldots, 1$. For $1 \le i \le n-1$, let $Y_i := y_i + y_{i+1}$ and define A_i to be the event that i is an ancestor of $i+1$. To compute $Q_T^k(A_i)$ note that record i is not an ancestor of $i+1$ in T. To become an ancestor by time k, record i must be requested. After its last request, record $i+1$ cannot be selected.

Conditioning on when record i is last requested, we find

$$Q_T^k(A_i) = y_i \sum_{j=0}^{k-1} (1-Y_i)^j = \frac{y_i}{Y_i}(1-(1-Y_i)^k).$$

Letting $k \to \infty$ shows that

$$Q^\infty(A_i) = y_i/Y_i. \tag{8}$$

Now let $N := \sum_{i \text{ odd}} I_{A_i}$, where I_A is the indicator of the event A. Then

$$E^\infty(N) = \sum_{i \text{ odd}} Q^\infty(A_i) = \sum_{i \text{ odd}} y_i/Y_i, \tag{9}$$

and

$$E_T^k(N) = \sum_{i \text{ odd}} \frac{y_i}{Y_i}(1-(1-Y_i)^k). \tag{10}$$

For $|i-j| \geq 2,\ Q^\infty(A_i \cap A_j) = Q^\infty(A_i)Q^\infty(A_j)$. Hence

$$\mathrm{Var}^\infty(N) = \sum_{i \text{ odd}} Q^\infty(A_i)(1-Q^\infty(A_i)) = \sum_{i \text{ odd}} \frac{y_i y_{i+1}}{Y_i^2} \leq \frac{n}{4}. \tag{11}$$

To compute $Q_T^k(A_i \cap A_j)$, we condition on the times records i and j are last requested:

$$\begin{aligned}
Q_T^k(A_i \cap A_j) &= y_i y_j \sum_{l=0}^{k-2} (1-Y_i)^l \sum_{m=0}^{k-2-l} (1-Y_i-Y_j)^m \\
&\quad + y_j y_i \sum_{l=0}^{k-2} (1-Y_j)^l \sum_{m=0}^{k-2-l} (1-Y_j-Y_i)^m \\
&= \frac{y_i y_j}{Y_i Y_j}\left[\frac{Y_j}{Y_i+Y_j} - (1-Y_i)^k + \frac{Y_i}{Y_i+Y_j}(1-Y_i-Y_j)^k\right] \\
&\quad + \frac{y_j y_i}{Y_j Y_i}\left[\frac{Y_i}{Y_j+Y_i} - (1-Y_j)^k + \frac{Y_j}{Y_j+Y_i}(1-Y_j-Y_i)^k\right] \\
&= \frac{y_i y_j}{Y_i Y_j}\left[1-(1-Y_i)^k-(1-Y_j)^k+(1-Y_i-Y_j)^k\right] \\
&\leq \frac{y_i y_j}{Y_i Y_j}\left[1-(1-Y_i)^k-(1-Y_j)^k+(1-Y_i)^k(1-Y_j)^k\right] \\
&= Q_T^k(A_i)Q_T^k(A_j).
\end{aligned}$$

Thus $\mathrm{Cov}_T^k(I_{A_i}, I_{A_j}) \le 0$ and

$$\begin{aligned}
\mathrm{Var}_T^k(N) &\le \sum_{i \text{ odd}} Q_T^k(A_i)(1 - Q_T^k(A_i)) \\
&= \sum_{i \text{ odd}} \left(\frac{y_i}{Y_i} - \frac{y_i}{Y_i}(1-Y_i)^k\right)\left(\frac{y_{i+1}}{Y_i} + \frac{y_i}{Y_i}(1-Y_i)^k\right) \\
&\le \sum_{i \text{ odd}} \left(\frac{y_i y_{i+1}}{Y_i^2} + \frac{y_i(y_i - y_{i+1})(1-Y_i)^k}{Y_i^2}\right) \\
&\le \sum_{i \text{ odd}} \left(\frac{y_i y_{i+1}}{Y_i^2} + \frac{y_i^2}{Y_i^2}(1-Y_i)^k\right).
\end{aligned}$$

We have

$$\begin{aligned}
\mathrm{Var}_T^k(N) - \mathrm{Var}^\infty(N) &\le \sum_{i \ odd} \frac{y_i^2}{Y_i^2}(1-Y_i)^k \\
&\le \sum_{i \text{ odd}} \frac{y_i}{Y_i}(1-Y_i)^k \\
&= E^\infty(N) - E_T^k(N) =: \Delta_k. \qquad (12)
\end{aligned}$$

Now let $B := \{N \le E^\infty(N) - \Delta_k/2\}$. Once we determine appropriate values of k we will show that the probability of B is small in stationarity and large until k steps of MTR.

From Chebychev's inequality,

$$Q^\infty(B) \le \frac{4\mathrm{Var}^\infty(N)}{\Delta_k^2} \qquad (13)$$

and

$$Q_T^k(B) = Q_T^k(N - E_T^k(N) \le \Delta_k/2) \ge 1 - \frac{4\mathrm{Var}_T^k(N)}{\Delta_k^2}. \qquad (14)$$

From (12),

$$\frac{\mathrm{Var}_T^k(N)}{\Delta_k^2} \le \frac{\mathrm{Var}^\infty(N) + \Delta_k}{\Delta_k^2} = \frac{\mathrm{Var}^\infty(N)}{\Delta_k^2} + \frac{1}{\sqrt{\mathrm{Var}^\infty(N)}}\left(\frac{\mathrm{Var}^\infty(N)}{\Delta_k^2}\right)^{1/2}.$$

Hence

$$\begin{aligned}
d(k) &\ge d(k;\mathbf{y}) \ge Q_T^k(B) - Q^\infty(B) \\
&\ge 1 - \frac{4}{\sqrt{\mathrm{Var}^\infty(N)}}\left(\frac{\mathrm{Var}^\infty(N)}{\Delta_k^2}\right)^{1/2} - \frac{8\mathrm{Var}^\infty(N)}{\Delta_k^2}. \qquad (15)
\end{aligned}$$

Now choose $\sigma = \text{rev}$, so that $y_1 \geq y_2 \geq \cdots \geq y_n$. Further, assume for definiteness that n is odd. We then have

$$\begin{aligned}
\text{Var}^{\infty}(N) &= \sum_{i \text{ odd}} \frac{y_i y_{i+1}}{(y_i + y_{i+1})^2} \\
&= \sum_{i \text{ odd}} \frac{p_{n+1-i}\, p_{n-i}}{(p_{n+1-i} + p_{n-i})^2} \\
&\geq \frac{1}{4} \sum_{i \text{ odd}} \left(\frac{p_{n-i}}{p_{n+1-i}} \right)^2 \geq \frac{n}{8} \min_{1 \leq i \leq n-1} \left(\frac{p_i}{p_{i+1}} \right)^2 . \qquad (16)
\end{aligned}$$

Also,

$$\begin{aligned}
\Delta_k &\geq \frac{1}{2} \sum_{i \text{ odd}} (1 - y_i - y_{i+1})^k \\
&= \frac{1}{2} \sum_{i \text{ odd}} (1 - p_{n+1-i} - p_{n-i})^k \\
&= \frac{1}{2} \sum_{j \text{ odd}} (1 - p_{j-1} - p_j)^k \geq \frac{1}{2} \sum_{j \text{ odd}} (1 - 2p_j)^k,
\end{aligned}$$

which, in combination with (11), gives

$$\frac{\text{Var}^{\infty}(N)}{\Delta_k^2} \leq \frac{n}{\left[\sum_{j \text{ odd}} (1 - 2p_j)^k \right]^2}. \qquad (17)$$

To make further progress in bounding (17) we need some conditions on the weights. The following theorem, following roughly along the lines of Theorem 2 in Diaconis (1993), gives a general result for obtaining a lower bound on variation distance.

Theorem 4 *For a triangular array of weights $0 < p_{n,1} \leq \cdots \leq p_{n,n}$ with $\sum_{j=1}^{n} p_{n,j} = 1$, suppose that the following conditions are satisfied:*

A1 $\max_{1 \leq j \leq n-1} \frac{p_{n,j+1}}{p_{n,j}} = O(n^{1/2})$.

A2 *There exists a function f with $f(n) \to \infty$ such that*

$$\frac{p_{n,j}}{p_{n,1}} \leq 2, \quad 2 \leq j \leq f(n).$$

A3 *There exists a function g with $g(n)/n^{1/2} \to \infty$ such that*

$$\frac{p_{n,j}}{p_{n,1}} = 1 + (1 + o(1)) \frac{j-1}{g(n)}, \quad \textit{uniformly in } 2 \leq j \leq f(n).$$

A4 $g(n)/(b(n,c)f(n)) = O(1)$ *and* $b(n,c)p_{n,1} = o(1)$ *as* $n \to \infty$, *where*

$$b(n,c) := \ln\left(\frac{g(n)}{\sqrt{n}}\right) - \ln\ln\left(\frac{g(n)}{\sqrt{n}}\right) + c.$$

Then $k(n,c) = b(n,c)/(2p_{n,1})$ *steps are necessary for convergence to stationarity for MTR.*

Remark:

For our weight classes it is easy to check whether these conditions are satisfied. Condition A1 is very weak and holds, even with $O(n^{1/2})$ reduced to $O(1)$, for all our weights. The condition A2 fails for geometric and power law weights, and condition A3 fails for uniform weights. All of the conditions are satisfied for generalized Zipf's law weights for any $s > 0$, with $f(n) = (1 - 2^{-1/s})n$ and $g(n) = n/s$. The result of Theorem 1 for generalized Zipf's law weights is then straightforward.

We now prove Theorem 4. As indicated in the above remark, this will conclude the proof for necessity in Theorem 1 for Zipf's law and GZL weights. After the proof of Theorem 4 we will treat the cases of uniform, power law, and geometric weights.

Proof For ease of notation we write p_j for $p_{n,j}$. Assume that conditions A1–A4 hold. Let $k = b/(2p_1)$.

From condition A1 and (16) it follows that there exists a constant $\delta > 0$ such that $\mathrm{Var}^\infty(N) \geq \delta$ for all n. Combining this with (15), we obtain

$$d(k) \geq 1 - 4(\delta^{-1/2} + 2)\left(\frac{\mathrm{Var}^\infty(N)}{\Delta_k^2}\right)^{1/2} \tag{18}$$

provided $\mathrm{Var}^\infty(N)/\Delta_k^2 \leq 1$. The remaining course of the proof is clear. We need to make $\mathrm{Var}^\infty(N)/\Delta_k^2$ suitably small (using (17)) in order to ensure that $d(k)$ is close to 1.

By condition A2, for $1 \leq j \leq f(n)$ we have $0 \leq 2p_j \leq 4p_1$, and $p_1 = o(1)$ follows (for example) from A3. Hence, uniformly for $1 \leq j \leq f(n)$,

$$-k\ln(1 - 2p_j) = k(2p_j + O(p_1^2)) = 2kp_j + O(bp_1) = 2kp_j + o(1),$$

where the last equality is from A4. This gives

$$\sum_{j \text{ odd}} (1-2p_j)^k \geq \sum_{\substack{1\leq j\leq f(n)\\ j \text{ odd}}} (1-2p_j)^k \geq (1+o(1)) \sum_{\substack{1\leq j\leq f(n)\\ j \text{ odd}}} e^{-2kp_j}.$$

At the same time, by A3

$$\begin{aligned}
\sum_{\substack{1\leq j\leq f(n)\\ j \text{ odd}}} e^{-2kp_j} &= \sum_{\substack{1\leq j\leq f(n)\\ j \text{ odd}}} \exp[-b(p_j/p_1)] \\
&= \sum_{\substack{1\leq j\leq f(n)\\ j \text{ odd}}} \exp\left\{-b\left[1+(1+o(1))\frac{j-1}{g(n)}\right]\right\} \\
&= \frac{e^{-b}\left[1-\exp\left\{-(1+o(1))bf(n)/g(n)\right\}\right]}{1-\exp\{-(1+o(1))2b/g(n)\}}. \qquad (19)
\end{aligned}$$

From A4 and (19) it follows that

$$\sum_{\substack{1\leq j\leq f(n)\\ j \text{ odd}}} e^{-2kp_j} \geq (1+o(1))e^{-b}\frac{g(n)}{2b}\epsilon \qquad (20)$$

for some positive constant ϵ. We conclude from (17) and (20) that

$$\begin{aligned}
\frac{\text{Var}^\infty(N)}{\Delta_k^2} &\leq (1+o(1))ne^{2b}\left(\frac{2b}{\epsilon g(n)}\right)^2 \\
&= (1+o(1))\left[2n^{1/2}e^b\frac{b}{\epsilon g(n)}\right]^2 \\
&= (1+o(1))\frac{4}{\epsilon^2}e^{2c},
\end{aligned}$$

the last equality holding by A4. ■

Uniform weights do not satisfy A3, but from (17),

$$\frac{\text{Var}^\infty(N)}{\Delta_k^2} \leq \frac{n}{\left[\frac{n}{2}(1-\frac{2}{n})^k\right]^2} = \frac{4}{n}\left(1-\frac{2}{n}\right)^{-2k}.$$

Setting $k = (n/4)(\ln n + c)$,

$$\frac{\text{Var}^\infty(N)}{\Delta_k^2} \leq e^c + o(1),$$

and thus $d(k) \geq 1 - 8e^c + o(1)$.

The proof of Theorem 1 for geometric weights follows from Theorem 3 in Diaconis (1993). For completeness we give essentially the same argument for power law weights.

Suppose record i is requested with probability $y_i = p_1$ and record $i+1$ is requested with probability $y_{i+1} = p_2$. Let $B = \{i <_a i+1\}$. As we have shown (recall (8)), $Q^\infty(B) = p_1/(p_1+p_2) \leq 1/2$. If we start from a tree in which i is an ancestor of $i+1$, then B necessarily occurs if $i+1$ is not requested. So

$$\begin{aligned} d(k) &\geq |Q^k(B) - Q^\infty(B)| \geq (1-p_2)^k - 1/2 \\ &\geq e^{-4k/\binom{n+1}{2}} - 1/2 = e^{-c} - 1/2 \to 1/2 \text{ as } c \searrow 0 \end{aligned}$$

for $k = \frac{c}{4}\binom{n+1}{2}$, where we have used the bound $1 - x \geq e^{-2x}$ for $0 \leq x \leq 1/2$.

Remarks:

1. In Diaconis's treatment for MTF, the "bad" event B is the event that record i is to the left of record $i+1$. Our event B is the lumped version of this in the context of trees.

2. In the case of uniform weights we can show that $k = (n/4)(\ln n - c)$ steps are necessary *regardless of the starting state.* For initial tree T define the event A_i by

$$A_i := \begin{cases} \{i \text{ is an ancestor of } i+1\}, & \text{if } i+1 <_a^T i \\ \{i+1 \text{ is an ancestor of } i\}, & \text{if } i <_a^T i+1. \end{cases}$$

With this event A_i the necessity part of the proof of Theorem 1 goes through similarly.

4 Expected search cost

4.1 General weights

In his Section 5, Fill (1993) obtains rates of convergence and explicit upper and lower bounds for the discrepancy between expected search cost at time k and in stationarity for uniform, geometric, and Zipf's law weights under MTF. There are two difficulties in extending these results to MTR: First,

expected search cost for MTR depends heavily on the order of the weight vector $\mathbf{y}$; second, there is no monotone relationship between the weights and the discrepancy at time k and in stationarity. These are in marked contrast to MTF for linear lists.

In this subsection we make some observations about the behavior of expected search cost for general weights. In the next subsection we analyze the case of uniform weights.

Given $T \in B_n$, the cost of accessing record i in T is equal to the number of comparisons used in searching for i in T. This is just one greater than the level of the node containing record i (where we define the level of the root node to be 0). Let $L_T(i)$ denote the level of record i in T. The average search cost for T, denoted $\mathrm{ASC}(T)$, is defined to be

$$\mathrm{ASC}(T) := \sum_{i=1}^{n} y_i L_T(i).$$

We denote by ESC^∞ the expected search cost over B_n in stationarity:

$$\mathrm{ESC}^\infty = E[\mathrm{ASC}(T)] \text{ with } T \sim Q^\infty.$$

Denote by ESC^k_π the expected search cost over B_n at time k when the initial distribution of the MTR chain is π. Allen and Munro (1978) show that

$$\mathrm{ESC}^\infty = 1 + 2\sum_{i<j} \frac{y_i y_j}{y_i + \ldots + y_j}. \tag{21}$$

For $i < j$, let $Y_{ij} := y_i + \ldots + y_j$. To calculate ESC^k_π, let $\mathbf{1}(A)$ denote the indicator of A and note that

$$L_T(i) = 1 + \sum_{j \neq i} \mathbf{1}(j <^T_a i),$$

$$Q^k_T(i <_a j) = \mathbf{1}(i <^T_a j)(1 - Y_{ij})^k + \frac{y_i}{Y_{ij}}(1 - (1 - Y_{ij})^k),$$

and

$$Q^k_T(j <_a i) = \mathbf{1}(j <^T_a i)(1 - Y_{ij})^k + \frac{y_j}{Y_{ij}}(1 - (1 - Y_{ij})^k).$$

We now have

$$\begin{aligned}
\mathrm{ESC}_\pi^k &= 1+\sum_{i=1}^n y_i E_\pi^k L(i) \\
&= 1+\sum_{T\in B_n}\pi(T)\sum_{i<j}\left[y_i Q_T^k(j <_a i)+y_j Q_T^k(i <_a j)\right] \qquad (22)\\
&= 1+\sum_{T\in B_n}\pi(T)\sum_{i<j}\Big[(y_i\mathbf{1}(j <_a^T i)+y_j\mathbf{1}(i <_a^T j))(1-Y_{ij})^k \\
&\quad +\frac{2y_iy_j}{Y_{ij}}(1-(1-Y_{ij})^k)\Big] \\
&= 1+2\sum_{i<j}\frac{y_iy_j}{Y_{ij}} \\
&\quad +\sum_{T\in B_n}\pi(T)\sum_{i<j}\left[\left(y_i\mathbf{1}(j <_a^T i)+y_j\mathbf{1}(i <_a^T j)-\frac{2y_iy_j}{Y_{ij}}\right)(1-Y_{ij})^k\right] \\
&= \mathrm{ESC}^\infty + D(\pi,k)
\end{aligned}$$

with

$$D(\pi,k) := \sum_{i<j}\left(y_iQ_\pi(j <_a i)+y_jQ_\pi(i <_a j)-\frac{2y_iy_j}{Y_{ij}}\right)(1-Y_{ij})^k.$$

Note that we recover (21) by letting $k\to\infty$.

We will focus on the case for which the initial state is a deterministic tree T, and write

$$D(T,k)=\sum_{i<j}\left(y_i\mathbf{1}(j <_a^T i)+y_j\mathbf{1}(i <_a^T j)-\frac{2y_iy_j}{Y_{ij}}\right)(1-Y_{ij})^k. \qquad (23)$$

Note that the quantities ESC^∞, ESC_T^k, and $D(T,k)$ depend on the ordering of the weights. We will sometimes write $\mathrm{ESC}^\infty(\mathbf{y})$, $\mathrm{ESC}_T^k(\mathbf{y})$, and $D(T,k;\mathbf{y})$ to indicate this dependence.

In studying convergence rates for expected search cost we would like to bound the relative error $D(T,k)/\mathrm{ESC}^\infty$ in a fashion similar to Fill (1993) for MTF. However, the dependence of (23) on both the initial tree and the permutation σ appears to make such a study, in general, intractable.

On the basis of intuition and Fill's results for MTF, it might seem a reasonable guess that $D(T,k)$ is largest when the initial tree is the degenerate

tree $t(\mathrm{id})$ (corresponding to the identity permutation, as described at the end of Section 2) and the weight vector is taken in increasing order, but this is not the case. For example, for $n = 3$ take $\mathbf{y}^{(1)} = (.1, .1, .8)$ and $\mathbf{y}^{(2)} = (.1, .8, .1)$. Then

$$\mathrm{ESC}^{\infty}(\mathbf{y}^{(1)}) = 1.43\overline{7} \geq 1.37\overline{5} = \mathrm{ESC}^{\infty}(\mathbf{y}^{(2)}).$$

Letting $T_1 = t((1, 2, 3))$ and $T_2 = t((1, 3, 2))$, one can show by direct calculation that $D(T_1, k; \mathbf{y}^{(1)}) \leq D(T_2, k; \mathbf{y}^{(2)})$ for all k and hence also that $D(T_1, k; \mathbf{y}^{(1)})/\mathrm{ESC}^{\infty}(\mathbf{y}^{(1)}) \leq D(T_2, k; \mathbf{y}^{(2)})/\mathrm{ESC}^{\infty}(\mathbf{y}^{(2)})$.

Numerical experiments with $n = 4$ show that $\min_{\sigma} \mathrm{ESC}^{\infty}(\mathbf{p} \circ \sigma)$ is not uniformly achieved for all $\mathbf{p}$ by any particular permutation σ.

Nor is it true that $\max_{\sigma} \mathrm{ESC}^{\infty}(\mathbf{p} \circ \sigma)$ is achieved by taking the weights to be equal. Letting $\mathbf{y} = (.255, .235, .245, .265)$ gives higher expected search cost than $\mathbf{y} = (.25, .25, .25, .25)$. Interestingly, numerical experiments suggest that $\max_{\sigma \in S_n} \mathrm{ESC}^{\infty}(\mathbf{p} \circ \sigma)$ is achieved at the two lists (p_3, p_1, p_2, p_4) and (p_4, p_2, p_1, p_3). However, we do not know how in general to characterize the ordering, if any, which maximizes $\mathrm{ESC}^{\infty}(\mathbf{y})$.

4.2 Uniform weights

In the case of general weights, the observations in Section 4.1 damper hopes of getting reasonable bounds for the relative error in approximating ESC^{∞} by ESC_T^k. In the case of equal weights, however, we show that $k = cn/\ln n$ steps are necessary and sufficient to drive the maximum relative error to 0.

Theorem 5 *For $T \in B_n$, let $E(T, k) := D(T, k)/\mathrm{ESC}^{\infty}$ be the relative error in approximating ESC^{∞} by ESC_T^k. Let $k = cn/\ln n$ with $c > 0$. If $p_i \equiv 1/n$, then for $n \geq 5$ sufficiently large that $c \geq (\ln n \ln 2)/(n - 4)$,*

$$\frac{1}{48c} e^{-\frac{6c}{\ln 5}} \leq \max_T E(T, k) \leq \frac{1}{c}.$$

Proof For equal weights it is clear from (23) that $D(T, k)$ is maximized by taking T to be any tree having the property that for all $i < j$ either $i <_a^T j$ or $j <_a^T i$. In particular,

$$\max_T D(T, k) \quad = \quad D(t(\mathrm{id}), k)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{1}{n}\left(1-\frac{2}{j-i+1}\right)\left(1-\frac{j-i+1}{n}\right)^k \\
&= \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=2}^{n-i+1}\left(1-\frac{2}{j}\right)\left(1-\frac{j}{n}\right)^k \\
&\leq \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=2}^{n-i+1} e^{-jk/n} \\
&\leq \frac{e^{-2k/n}}{1-e^{-k/n}}.
\end{aligned}
$$

From (21) we have

$$\mathrm{ESC}^{\infty} = 2\left(\frac{n+1}{n}\right)H_n - 3 \geq \ln n, \tag{24}$$

where $H_n := \sum_{i=1}^{n} i^{-1}$ is the nth harmonic number. Thus

$$\max_T E(T,k) = E(t(\mathrm{id}),k) \leq \frac{e^{-2k/n}}{(1-e^{-k/n})\ln n}.$$

It is straightforward to show that $e^{-2x}/(1-e^{-x}) \leq 1/x$ for $x > 0$. Thus for $c > 0$, taking $k = cn/\ln n$ gives

$$\max_T E(T,k) \leq \frac{1}{c}.$$

For the lower bound, assume, for simplicity, that n is even. Then

$$
\begin{aligned}
D(t(\mathrm{id}),k) &\geq \frac{1}{n}\sum_{i=1}^{n/2}\sum_{j=2}^{n/2}\left(1-\frac{2}{j}\right)\left(1-\frac{j}{n}\right)^k \\
&\geq \frac{1}{2}\sum_{j=2}^{n/2}\left(1-\frac{2}{j}\right)e^{-2jk/n} \\
&\geq \frac{1}{6}\sum_{j=3}^{n/2} e^{-2jk/n} \\
&\geq \frac{n}{12k}(e^{-6k/n} - e^{-k}e^{-2k/n}) \\
&\geq \frac{n}{24k}e^{-6k/n}
\end{aligned}
$$

for $n \geq 5$ and $k \geq (n \ln 2)/(n-4)$, where we have used the bound $1-x \geq e^{-2x}$ for $0 \leq x \leq 1/2$. We have $\mathrm{ESC}^{\infty} \leq 2 \ln n$ for $n \geq 3$; taking $k = cn/\ln n$ gives

$$\begin{aligned} \max_T E(T,k) &\geq \frac{1}{48c} e^{-6c/\ln n} \\ &\geq \frac{1}{48c} e^{-6c/\ln 5} \end{aligned}$$

for $n \geq 5$ and $c \geq (\ln n \ln 2)/(n-4)$.

■

Remarks:

1. Asymptotically, the only lack of sharpness in the upper bound of Theorem 5 comes from (24). Indeed, it is not hard to show that

$$\lim_{n\to\infty} \max_T E(T, cn/\ln n) = \frac{1}{2c}.$$

2. A *perfect tree* on $n = 2^m - 1$ nodes is a binary search tree in which all leaves have the same depth. A *complete tree* of height h is formed from a perfect tree of height $h-1$ by adding one or more leaves at depth h; these leaves must be filled in at the leftmost available positions. We can show for uniform weights that the complete tree (call it T_n) on n nodes minimizes $D(T,k)$ of (23) over $T \in B_n$ for every $k \geq 0$. Furthermore, for $n = 2^m - 1$ we have

$$D(T_n,k) = -\left(1+\frac{1}{n}\right) \sum_{j=2}^{n} \left(2j^{-1} - 2^{-\lfloor \log_2(j-1) \rfloor}\right) \left(1-\frac{j}{n}\right)^k,$$

from which one sees that $k = n^{1-1/c}$ steps are sufficient to make

$$\max_T [-E(T,k)] = -E(T_n,k) > 0$$

small. We have not investigated necessity, but (for $n = 2^m - 1$) this result shows that Theorem 5 gives a worst-case rate for convergence (measured by absolute value of relative error) for expected search cost.

3. For MTR with equal weights, expected search cost converges to stationarity much faster that does the distribution of the tree, which takes

of order $n \ln n$ steps. This discrepancy in rates was observed by Fill (1993) for the linear list case for uniform, Zipf's law, and geometric weights. As for those examples for MTF, expected search cost for MTR using uniform weights exhibits no cutoff phenomenon.

5 References

Aldous, D. and Diaconis, P. (1986). Shuffling cards and stopping times. *Amer. Math. Monthly* **93** 333–348.

Aldous, D. and Diaconis, P. (1987). Strong uniform times and finite random walks. *Adv. in Appl. Math.* **8** 69–97.

Allen, B. and Munro, I. (1978). Self-organizing binary search trees. *J. ACM* **25** 526–535.

Bitner, J. R. (1979). Heuristics that dynamically organize data structures. *SIAM J. Comput.* **8** 82–110.

Diaconis, P. (1993). Analysis of some weighted mixing schemes. Unpublished manuscript.

Diaconis, P. and Fill, J. A. (1990). Strong stationary times via a new form of duality. *Ann. Prob.* **18** 1483–1522.

Dobrow, R. P. and Fill, J. A. (1993). On the Markov chain for the move-to-root rule for binary search trees. Technical Report #530, Department of Mathematical Sciences, The Johns Hopkins University.

Fill, J. A. (1993). An exact formula for the move-to-front rule for self-organizing lists. Technical Report #529, Department of Mathematical Sciences, The Johns Hopkins University.

Hendricks, W. J. (1989). *Self-organizing Markov Chains.* MITRE Corp., McLean, Va.

Knuth, D. E. (1973). *The Art of Computer Programming.* Vol. 3. Addison–Wesley, Reading, Mass.

Rivest, R. (1976). On self-organizing sequential search heuristics. *Comm. ACM* **19** 63–67.

Schwartz, E. S. (1963). A dictionary for minimum redundancy encoding. *J. ACM* **10** 413–439.

Robert P. Dobrow
Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD 21218-2689

James Allen Fill
Department of Mathematical Sciences
The Johns Hopkins University
Baltimore, MD 21218-2689