

## RATES OF CONVERGENCE FOR THE MOVE-TO-ROOT MARKOV CHAIN FOR BINARY SEARCH TREES<sup>1</sup>

BY ROBERT P. DOBROW<sup>2</sup> AND JAMES ALLEN FILL

*Johns Hopkins University*

The move-to-root heuristic is a self-organizing rule that 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 the maximum relative error to 0.

**1. Introduction and summary.** The move-to-root (MTR) self-organizing scheme for binary search trees has been studied in Dobrow and Fill (1995). They show that the Markov chain for MTR can be derived by lumping the Markov chain for move-to-front (MTF) on lists and determine numerous characteristics of the MTR chain, including the spectral decomposition of the transition matrix and formulas for  $k$ -step transition probabilities.

Here we treat rates of convergence to stationarity. From lumpability it is immediate that MTR converges to its stationary distribution at least as fast as does MTF. In this paper we are able to quantify the speedup under quite general circumstances.

We refer the reader to Dobrow and Fill (1995) for background on binary search trees, self-organizing data structures in general and self-organizing trees in particular, and the move-to-root heuristic.

In Section 2 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  $k/2$ .

---

Received November 1993; revised May 1994.

<sup>1</sup>Research for both authors supported by NSF Grant DMS-93-11367.

<sup>2</sup>Work performed while a doctoral student in the Department of Mathematical Sciences, Johns Hopkins University.

AMS 1991 subject classifications: Primary 60J10; secondary 68P10, 68P05.

Key words and phrases. Markov chains, convergence to stationarity, self-organizing search, binary search trees, move-to-root rule, move-to-front rule.

In Section 3 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. Convergence rates for MTR.** Let  $0 < p_1 \leq p_2 \leq \dots \leq p_n$  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  $Q^\infty$  denote the stationary distribution of MTR. Note that for fixed  $\mathbf{p}$  the quantities  $Q_T^k$  and  $Q^\infty$  depend on the permutation  $\sigma$ .

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  $T$ :

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

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  $\mathbf{p}_n = (p_{n,i}, i = 1, \dots, n)$ ,  $n \geq 1$ , say that  $k = k(n, c)$  steps are *sufficient* for total variation convergence to stationarity if there exist positive constants  $\alpha$  and  $\beta$  such that for each fixed  $c$  we have  $d(k) \leq \alpha e^{-\beta c} + o(1)$  as  $n \rightarrow \infty$ . 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 in succeeding text), we will even be able to take  $h(c) \rightarrow 1$  as  $c \rightarrow -\infty$ . 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 definition, 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  $i$ th 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)	$1/(n - i + 1)^s$ , $s > 0$ fixed
Power law	$i^s$ , $s > 0$ fixed
Geometric	$\theta^{n-i}$ , $0 < \theta < 1$ fixed

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	$c\theta^{-n}$	$c\theta^{-n}$

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 gives rates of convergence to stationarity for MTR.*

**REMARKS.**

1. Because 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, for example, 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 (1995) along with the observation that  $\Pi(T)$  is a singleton when  $T$  is the tree corresponding to either of the permutations  $\text{id} = (1, \dots, n)$  or  $\text{rev} = (n, \dots, 1)$ . 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.

**2.1. Proof of Theorem 1: Sufficiency.** Coupling is a probabilistic technique that 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,

$$(1) \quad d(k; \mathbf{y}) \leq P(T > k), \quad k \geq 0.$$

Before describing the coupling we introduce some terminology following Dobrow and Fill (1995). For  $R \subseteq [n]$ , write  $r_1 < r_2 < \dots < r_m$  for the elements of  $R$ . Define  $r_0 := 0$  and  $r_{m+1} := n + 1$ . Let

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

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*

$$(2) \quad d(k; \mathbf{y}) \leq P(T > k) \leq \sum_{i=1}^{n-1} (1 - y_i - y_{i+1})^k.$$

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 2.1.** *Let  $i \in [n]$  and let  $r_1, r_2, \dots$  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  $i <_a^{S^{k'}} j$ .
- (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 3.1 in Dobrow and Fill (1995).  $\square$

We now prove Theorem 2.

**PROOF OF THEOREM 2.** 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 (1995), 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  $R = \{r_1, \dots, r_T\}$  be the set of records requested through time  $T$ .

If  $i \in R$ , then  $i <_a^{Y^T} j$  by Lemma 2.1(a). If  $i \notin R$ , then  $i + 1 \in R$  because  $R$  has no gaps of size greater than 1. If  $j \in R$ , then  $i <_a^{Y^T} j$  by Lemma 2.1. If  $j \notin R$ , we will derive a contradiction. Because  $j \notin R$  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  $i \not<_a^{X^{k'}} j$ . Because  $i \notin R$ , 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 (2) follows because  $T$  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

$$(3) \quad P(T > k) = P\left(\bigcup_{i=1}^{n-1} A_i\right)$$

and the second inequality in (2) follows from subadditivity.  $\square$

**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  $\sigma$  of the weights.
2. Our coupling time is not a fastest coupling time, for which the inequality in (1) would be an equality. For example, stopping when both trees agree is faster (but still not fastest).

**THEOREM 3.** For weights  $0 < p_1 \leq \dots \leq p_n$ , suppose there exist positive constants  $\alpha$  and  $\beta$  such that

$$(4) \quad \sum_{i=1}^{n-1} (1 - p_i)^{2k} \leq \alpha e^{-\beta c}.$$

Then  $k$  steps are sufficient for MTR.

PROOF. Recalling the notation set forth at the beginning of Section 2, 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.  $\square$

For a fixed class of weights, let  $k$  be the number in the sufficiency column of Table 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  $\alpha$  and  $\beta$ . For uniform, Zipf’s law and GZL weights, Diaconis (1993) showed that

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

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

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

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 does not 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 (5) is satisfied and  $kp_1 \rightarrow \infty$ . 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 (2). 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

$$(6) \quad P(T > k) = \sum_{u=1}^n P_k(u)P(E_u),$$

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, whereas  $P_k(u)$  is well known from ordinary coupon collecting. We have made a careful asymptotic analysis of (6) and find that the subadditive bound on (3) gives the correct lead order term.

*2.2. 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  $\text{rev}$  is the reversal permutation. That is,  $T$  has record  $n$  at the root, and record  $i + 1$  is the parent of record  $i$  for  $i = n - 1, \dots, 1$ . For  $1 \leq i \leq n - 1$ , 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  $k \rightarrow \infty$  shows that

$$(7) \quad Q^\infty(A_i) = y_i/Y_i.$$

Now let  $N := \sum_{i \text{ odd}} \mathbf{1}(A_i)$ , where  $\mathbf{1}(A)$  is the indicator of the event  $A$ . Then

$$(8) \quad E^\infty(N) = \sum_{i \text{ odd}} Q^\infty(A_i) = \sum_{i \text{ odd}} y_i/Y_i,$$

and

$$(9) \quad E_T^k(N) = \sum_{i \text{ odd}} \frac{y_i}{Y_i} (1 - (1 - Y_i)^k).$$

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

$$(10) \quad \text{Var}^\infty(N) = \sum_{i \text{ odd}} Q^\infty(A_i)(1 - Q^\infty(A_i)) = \sum_{i \text{ odd}} \frac{y_i y_{i+1}}{Y_i^2} \leq \frac{n}{4}.$$

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

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

Thus  $\text{Cov}_T^k(\mathbf{1}(A_i), \mathbf{1}(A_j)) \leq 0$  and

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



We have

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

Now let  $B := \{N \leq E^\infty(N) - \Delta_k/2\}$ . 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,

$$(12) \qquad \qquad \qquad Q^\infty(B) \leq \frac{4 \text{Var}^\infty(N)}{\Delta_k^2}$$

and

$$(13) \qquad \qquad Q_T^k(B) = Q_T^k(N - E_T^k(N) \leq \Delta_k/2) \geq 1 - \frac{4 \text{Var}_T^k(N)}{\Delta_k^2}.$$

From (11),

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

Hence

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

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

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

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 (10), gives

$$(16) \quad \frac{\text{Var}^\infty(N)}{\Delta_k^2} \leq \frac{n}{\left[\sum_{j \text{ odd}} (1 - 2p_j)^k\right]^2}.$$

To make further progress in bounding (16), 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  $0 < p_{n,1} \leq \dots \leq p_{n,n}$  with  $\sum_{j=1}^n p_{n,j} = 1$ , suppose that the following conditions are satisfied:*

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

**A2.** *There exists a function  $f$  with  $f(n) \rightarrow \infty$  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  $g(n)/n^{1/2} \rightarrow \infty$  such that*

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

**A4.**  $g(n)/(b(n,c)f(n)) = O(1)$  and  $b(n,c)p_{n,1} = o(1)$  as  $n \rightarrow \infty$ , 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 foregoing 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 OF THEOREM 4. 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 (15) it follows that there exists a constant  $\delta > 0$  such that  $\text{Var}^\infty(N) \geq \delta$  for all  $n$ . Combining this with (14), we obtain

$$(17) \quad d(k) \geq 1 - 4(\delta^{-1/2} + 2) \left( \frac{\text{Var}^\infty(N)}{\Delta_k^2} \right)^{1/2}$$

provided  $\text{Var}^\infty(N)/\Delta_k^2 \leq 1$ . The remaining course of the proof is clear. We need to make  $\text{Var}^\infty(N)/\Delta_k^2$  suitably small [using (16)] 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

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

At the same time, by A3,

$$\begin{aligned} \sum_{\substack{1 \leq j \leq f(n) \\ j \text{ odd}}} \exp(-2kp_j) &= \sum_{\substack{1 \leq j \leq f(n) \\ j \text{ odd}}} \exp\left[-b\left(\frac{p_j}{p_1}\right)\right] \\ (18) \quad &= \sum_{\substack{1 \leq j \leq f(n) \\ j \text{ odd}}} \exp\left\{-b\left[1 + (1 + o(1))\frac{j-1}{g(n)}\right]\right\} \\ &= \frac{\exp(-b)[1 - \exp\{-(1 + o(1))bf(n)/g(n)\}]}{1 - \exp\{-(1 + o(1))2b/g(n)\}}. \end{aligned}$$

From A4 and (18) it follows that

$$(19) \quad \sum_{\substack{1 \leq j \leq f(n) \\ j \text{ odd}}} \exp(-2kp_j) \geq (1 + o(1)) \exp(-b) \frac{g(n)}{2b} \varepsilon$$

for some positive constant  $\varepsilon$ . We conclude from (16) and (19) that

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

the last equality holding by A4.  $\square$

Uniform weights do not satisfy A3, but from (16),

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

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 (7)],  $Q^\infty(B) = p_1/(p_1 + p_2) \leq 1/2$ . 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

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

for  $k = (c/4)\binom{n+1}{2}$ , 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.

### 3. Expected search cost.

**3.1. General weights.** In his Section 5, Fill (1995) 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 1 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  $\text{ASC}(T)$ , is defined to be

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

We denote by  $\text{ESC}^\infty$  the expected search cost over  $B_n$  in stationarity:

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

Denote by  $\text{ESC}_\pi^k$  the expected search cost over  $B_n$  at time  $k$  when the initial distribution of the MTR chain is  $\pi$ . Allen and Munro (1978) show that

$$(20) \quad \text{ESC}^\infty = 1 + 2 \sum_{i < j} \frac{y_i y_j}{y_i + \cdots + y_j}.$$

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

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

$$Q_T^k(i <_a j) = \mathbf{1}(i <_a^T j)(1 - Y_{ij})^k + \frac{y_i}{Y_{ij}}(1 - (1 - Y_{ij})^k)$$

and

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

We now have

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

with

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

Note that we recover (20) by letting  $k \rightarrow \infty$ .

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

$$(22) \quad D(T, k) = \sum_{i < j} \left( y_i \mathbf{1}(j <_a^T i) + y_j \mathbf{1}(i <_a^T j) - \frac{2y_i y_j}{Y_{ij}} \right) (1 - Y_{ij})^k.$$

Note that the quantities  $\text{ESC}^\infty$ ,  $\text{ESC}_T^k$  and  $D(T, k)$  depend on the ordering of the weights. We will sometimes write  $\text{ESC}^\infty(\mathbf{y})$ ,  $\text{ESC}_T^k(\mathbf{y})$  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  $D(T, k)/\text{ESC}^\infty$  in a fashion similar to Fill (1995) for MTF. However, the dependence of (22) on both the initial tree and the permutation  $\sigma$  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  $t(\text{id})$  (corresponding to the identity permutation) and the weight vector is taken in increasing order, but this is not the case. For example, for  $n = 3$

take  $\mathbf{y}^{(1)} = (0.1, 0.1, 0.8)$  and  $\mathbf{y}^{(2)} = (0.1, 0.8, 0.1)$ . Then

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

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  $D(T_1, k; \mathbf{y}^{(1)}) / \text{ESC}^\infty(\mathbf{y}^{(1)}) \leq D(T_2, k; \mathbf{y}^{(2)}) / \text{ESC}^\infty(\mathbf{y}^{(2)})$ .

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

Nor is it true that  $\max_\sigma \text{ESC}^\infty(\mathbf{p} \circ \sigma)$  is achieved by taking the weights to be equal. Letting  $\mathbf{y} = (0.255, 0.235, 0.245, 0.265)$  gives higher expected search cost than  $\mathbf{y} = (0.25, 0.25, 0.25, 0.25)$ . Interestingly, numerical experiments suggest that  $\max_{\sigma \in S_n} \text{ESC}^\infty(\mathbf{p} \circ \sigma)$  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  $\text{ESC}^\infty(\mathbf{y})$ .

**3.2. Uniform weights.** In the case of general weights, the observations in Section 3.1 dampen hopes of getting reasonable bounds for the relative error in approximating  $\text{ESC}^\infty$  by  $\text{ESC}_T^k$ . 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  $E(T, k) := D(T, k) / \text{ESC}^\infty$  be the relative error in approximating  $\text{ESC}^\infty$  by  $\text{ESC}_T^k$ . Let  $k = cn / \ln n$  with  $c > 0$ . If  $p_i \equiv 1/n$ , then for  $n \geq 5$  sufficiently large that  $c \geq (\ln 2)(\ln n) / (n - 4)$ ,*

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

**PROOF.** For equal weights, it is clear from (22) 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,

$$\begin{aligned} \max_T D(T, k) &= D(t(\text{id}), k) \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{1}{n} \left(1 - \frac{2}{j-i+1}\right) \left(1 - \frac{j-i+1}{n}\right)^k \\ &= \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=2}^{n-i+1} \left(1 - \frac{2}{j}\right) \left(1 - \frac{j}{n}\right)^k \\ &\leq \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=2}^{n-i+1} \exp(-jk/n) \\ &\leq \frac{\exp(-2k/n)}{1 - \exp(-k/n)}. \end{aligned}$$

From (20) we have

$$(23) \quad \text{ESC}^\infty = 2 \left( \frac{n+1}{n} \right) H_n - 3 \geq \ln n,$$

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

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

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

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

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  $\text{ESC}^\infty \leq 2 \ln n$  for  $n \geq 3$ . Taking  $k = cn/\ln n$  gives

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

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

REMARKS.

1. Asymptotically, the only lack of sharpness in the upper bound of Theorem 5 comes from (23). Indeed, it is not hard to show that

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



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 (22) over  $T \in B_n$  for every  $k \geq 0$ . Furthermore, for  $n = 2^m - 1$  we have

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

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 than does the distribution of the tree, which takes of order  $n \ln n$  steps. This discrepancy in rates was observed by Fill (1995) 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.

## 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. Assoc. Comput. Mach.* **25** 526–535.
- 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. Probab.* **18** 1483–1522.
- DOBROW, R. P. and FILL, J. A. (1995). On the Markov chain for the move-to-root rule for binary search trees. *Ann. Appl. Probab.* **5** 1–19.
- FILL, J. A. (1995). An exact formula for the move-to-front rule for self-organizing lists. *J. Theoret. Probab.* To appear.
- KNUTH, D. E. (1973). *The Art of Computer Programming* **3**. Addison-Wesley, Reading, MA.
- SCHWARTZ, E. S. (1963). A dictionary for minimum redundancy encoding. *J. Assoc. Comput. Mach.* **10** 413–439.

NATIONAL INSTITUTE OF  
STANDARDS AND TECHNOLOGY  
STATISTICAL ENGINEERING DIVISION  
ACADEMIC BUILDING, A337  
GAITHERSBURG, MARYLAND 20899-0001  
dobrow@cam.nist.gov

DEPARTMENT OF MATHEMATICAL SCIENCES  
JOHNS HOPKINS UNIVERSITY  
BALTIMORE, MARYLAND 21218-2689  
jimfill@jhvm.hcf.jhu.edu