Operations Research Letters 25 (1999) 109–118

operations research letters

www.elsevier.com/locate/orms

Performance of the move-to-front algorithm with Markov-modulated request sequences

E.G. Coffman Jr.[a], Predrag Jelenković[b,*]

[a] *Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, NJ 07974, USA*
[b] *Department of Electrical Engineering, Columbia University, 500 West 120th Street, New York, NY 10027, USA*

Received 1 October 1998; received in revised form 1 March 1999

Abstract

We study the classical move-to-front (MTF) algorithm for self-organizing lists within the Markov-modulated request (MMR) model. Such models are useful when list accesses are generated within a relatively small set of modes, with the request sequences in each mode being i.i.d. These modes are often called localities of reference and are known to exist in such applications as traffic streams of Ethernet or ATM networks and the locus of control or data accesses of executing computer programs. Our main results are explicit formulas for the mean and variance of the search-cost, the number of accesses required to find a given list element. By adjusting the number of modes, one can use the MMR methodology to trade off the quality of an approximation with the computational effort it requires. Thus, our results provide a useful new tool for evaluating the MTF rule in linear-search applications with correlated request sequences. We illustrate the computations with several examples. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Self-organizing lists; Move-to-front algorithm; Markov-modulated process; Hidden Markov-chains; Locality of reference; Internet modeling

1. Introduction

Performance analysis of self-organizing data structures, e.g., lists and trees, has a long history; references to the early work in this area can be found in [4,9]. Quite recently, interest in these data structures has been rekindled by cache design problem in modern distributed networks. One of the most popular heuristics for self-organizing lists has been the *move-to-front* (MTF) rule. MTF is defined on sequences of requests for elements in a given list of N elements. In processing each new request, MTF moves the requested element to the first position (i.e., the left-most position or head) of the list, if it is not already there; the ordering of the remaining $N-1$ elements of the list remains unchanged. If the new request is in the rth position, the

[*] Corresponding author.
E-mail addresses: egc@bell-labs.com (E.G. Coffman Jr.), predrag@ee.columbia.edu (P. Jelenković)

PII: S0167-6377(99)00037-1

cost of processing the request is r; this represents the number of comparisons needed to find the request in a linear search of the list.

In the analysis of self-organizing lists, there have been two approaches: probabilistic and combinatorial (amortized) analysis. We are interested in typical or average-case behavior, so we concentrate on the former approach. For the latter, the reader is referred to Bentley and McGeoch [1]. Our specific objective is a probabilistic analysis of the most common cost function: the position (search distance) of the currently requested element. The novelty of our contribution lies in our extension of MTF theory to Markov-modulated request (MMR) sequences.

There is a large literature on MMR models that spans a wide variety of applications in applied probability and engineering, including speech recognition, communications engineering, statistics, and risk theory. See [6] for many references. Equivalent names under which Markov-modulated models are known in the literature are: Markov-modulated random walks, random walks on Markov chains, functions (or random functions) of Markov chains, and hidden Markov chains. We define them formally as follows. Let $J=\{J_t,\, t=0,1,\dots\}$ be a discrete time, aperiodic, irreducible, finite Markov chain with state space $\{1,\dots,K\}$, and transition matrix $P=\{p_{ij}\}$, and let $R=\{R_t,\, t=0,1,\dots\}$ be a discrete-time process with state space $\{1,\dots,N\}$. We say that R is Markov-modulated if the pair process $\{(R_t,J_t),t\geqslant 0\}$ is Markovian in its second coordinate, i.e., if

$$\mathbf{P}[R_t=r,\, J_t=j|(R_{t-1},J_{t-1}=i),\dots,(R_0,J_0)]=\mathbf{P}[R_t=r,\, J_t=j|J_{t-1}=i]=p_{ij}q_{jr},$$

where $q_{jr}:=\mathbf{P}[R_t=r|J_t=j]$. To avoid trivialities we assume that for each r there exists at least one j such that $q_{jr}>0$. Define the stationary marginal distribution of R as $q_r:=\mathbf{P}[R_t=r]=\sum_{i=1}^{K}\pi_i q_{ir}$, where $\{\pi_i, 1\leqslant i\leqslant K\}$, is the stationary distribution of J. Since the process J, to be called the *modulating process*, is ergodic, its stationary distribution is unique and positive. We assume that J is in its stationary regime, i.e., the distribution of J_0 is chosen according to π. The process R models sequences of requests for elements of the set $L=\{1,\dots,N\}$. A discrete-time process $\{\sigma_t, t=0,1,\dots\}$ is induced by R and that MTF rule, where $\sigma_t=(\sigma_t(1),\dots,\sigma_t(N))$ is a list (permutation) of the elements of L. We assume that, at $t=0$, all permutations are equally likely, i.e., σ_0 is uniformly distributed over the set of all $N!$ permutations. According to MTF, σ_{t+1} is constructed from σ_t by bringing the element R_t to the first position of the list, if it is not already there, and keeping the ordering of the remaining elements unchanged; thus, the positions of the elements that were ahead of R_t are increased by one, while the positions of those behind R_t remain unchanged. Note that the joint process $\{(J_{t-1},\sigma_t),\, t\geqslant 1\}$ is a Markov chain.

Early work on the probabilistic analysis of MTF dealt with i.i.d. requests; see [7] for key references. Our MTF model lies within the broader framework of Markovian request sequences. The MTF scheme with time-dependent Markovian requests was investigated in [10], where a formula for the expected search cost was derived. Special cases of the Markov model that are analytically more tractable were investigated by Rodrigues [13], who also examined convergence to stationarity [12]. Dobrow and Fill [5] derived transient and stationary probabilities for MTF in the Markov model. They also investigated spectral properties and the rate of convergence to stationarity.

Our interest in the MMR model stemmed from its flexibility as a tractable model of the high-correlation (locality) structure encountered in the traffic streams (e.g., voice, video, and multimedia) of modern communication networks. (References to these models can be found in [7].) But more generally, our methods extend the computational tools for evaluating the performance of linear-search heuristics in an environment of Markovian request sources. They will be significantly more efficient than existing techniques when MMR models apply with the number K of modes relatively small.

The analysis of the MTF algorithm appears in Section 2, where we develop explicit formulas for the transient and steady-state mean and variance of the cost function. In Section 3 we give numerical examples of MMR models and study the performance of MTF in these models. We conclude in Section 4 with an application of our results to LRU caching.

2. Mean and variance of the search cost

We derive explicit formulas for the mean and variance of the search cost C_t: the current position of the element R_t in the list state σ_t. Thus, the search cost at time t is a measure of the time required by a linear search of the list to find the element requested at time t. The reversed Markov chain $\{\tilde{J}_t\}$ and the corresponding modulation process $\{\tilde{R}_t\}$ arise naturally in the following analysis. The transition probability matrix of $\tilde{J}$ is denoted by $\tilde{P} = \{\tilde{p}_{ij}\}$, where $\tilde{p}_{ij} = \pi_j p_{ji}/\pi_i$. For each $i \neq j$, let A_t^{ji} be the event that j is to the left of i in σ_t. Given that $R_t = r$, the cost C_t is the number of elements to the left of r in the list plus 1 to account for r itself. Thus, if we let $\sum^*_{r_1,\dots,r_k}$ denote the sum over all sequences of *distinct* elements $r_1,\dots,r_k$ with $r_i \in \{1,\dots,N\}$, $1 \leqslant i \leqslant k$, then we have the well-known indicator-function representation,

$$C_t = 1 + \sum_{k,r}^{*} \mathbf{1}(R_t = r, A_t^{kr})$$
$$= 1 + \sum_{1\leqslant i \leqslant K} \sum_{k,r}^{*} \mathbf{1}(J_t = i, R_t = r, A_t^{kr}) \tag{1}$$

and hence, after taking expectations,

$$\mathbf{E}C_t = 1 + \sum_{1\leqslant i \leqslant K} \sum_{k,r}^{*} \mathbf{P}[J_t = i, R_t = r, A_t^{kr}]. \tag{2}$$

The ith summand $\mathbf{P}[J_t = i, R_t = r, A_t^{kr}]$ in the above expression can be written

$$\mathbf{P}[J_t = i, R_t = r, A_t^{kr}] = \sum_{m=1}^{t} \mathbf{P}[J_t = i, R_t = r, R_{t-1} \notin \{r,k\},\dots, R_{t-m+1} \notin \{r,k\}, R_{t-m} = k]$$
$$+\mathbf{P}[J_t = i, R_t = r, R_{t-1} \notin \{r,k\},\dots, R_0 \notin \{r,k\}, A_0^{kr}].$$

Recall that σ_0 is a permutation chosen uniformly at random, so we have $\mathbf{P}[A_0^{kr}] = 1/2$ for all k, r. In terms of the reversed processes, we can write

$$\mathbf{P}[J_t = i, R_t = r, A_t^{kr}] = \sum_{m=1}^{t} \mathbf{P}[\tilde{J}_0 = i, \tilde{R}_0 = r, \tilde{R}_1 \notin \{r,k\},\dots, \tilde{R}_{m-1} \notin \{r,k\}, \tilde{R}_m = k]$$
$$+\frac{1}{2}\mathbf{P}[\tilde{J}_0 = i, \tilde{R}_0 = r, \tilde{R}_1 \notin \{r,k\},\dots, \tilde{R}_t \notin \{r,k\}]. \tag{3}$$

For a more compact notation, we introduce the matrices

$$Q_{r_1,\dots,r_k} = \{\tilde{p}_{ij}(q_{jr_1} + \cdots + q_{jr_k})\}, \quad \hat{Q}_{r_1,\dots,r_k} = \{\tilde{p}_{ij}(1 - q_{jr_1} - \cdots - q_{jr_k})\}. \tag{4}$$

Let $\rho(Q)$ denote the spectral radius of matrix Q. Then, by the assumptions on P and q_{jr} from the introduction and Corollary 1, p. 8 of [14], it follows that

$$\rho(Q_{r_1,\dots,r_k}) < 1 \tag{5}$$

for any choice of $r_1,\dots,r_k$. Next, the mth summand in (3) can be expressed as

$$\sum_{i_1,\dots,i_m} \mathbf{P}[\tilde{J}_0 = i, \tilde{R}_0 = r, \tilde{J}_1 = i_1, \tilde{R}_1 \notin \{r,k\},\dots, \tilde{J}_{m-1} = i_{m-1}, \tilde{R}_{m-1} \notin \{r,k\}, \tilde{J}_m = i_m, \tilde{R}_m = k]$$
$$= \sum_{i_1,\dots,i_m} \pi_i q_{ir} \tilde{p}_{ii_1}(1 - q_{i_1 r} - q_{i_1 k}) \cdots \tilde{p}_{i_{m-2}i_{m-1}}(1 - q_{i_{m-1}r} - q_{i_{m-1}k}) \tilde{p}_{i_{m-1}i_m} q_{i_m k}$$
$$= \pi_i q_{ir}(\hat{Q}_{rk}^{m-1} Q_k \mathbf{e})(i), \tag{6}$$

where $\boldsymbol{e}$ is a column vector of ones, and $(\hat{Q}_{rk}^{m-1}Q_k\boldsymbol{e})(i)$ is the ith element of the vector $\hat{Q}_{rk}^{m-1}Q_k\boldsymbol{e}$. Similarly, we find that

$$\mathbf{P}[\tilde{J}_0 = i, \tilde{R}_0 = r, \tilde{R}_1 \notin \{r,k\}, \dots, \tilde{R}_t \notin \{r,k\}] = \pi_i q_{ir}(\hat{Q}_{rk}^{t}\boldsymbol{e})(i). \tag{7}$$

Finally, after substituting (3), (6), and (7) into (2) and using (5), we arrive at the following result.

Theorem 1. *The expected search cost is expressed by*

$$\mathbf{E}C_t = 1 + \sum_{k,r}^{*}\left[\boldsymbol{v}_r(I-\hat{Q}_{rk})^{-1}Q_k\boldsymbol{e} + \boldsymbol{v}_r\hat{Q}_{rk}^{t}\left(-(I-\hat{Q}_{rk})^{-1}Q_k\boldsymbol{e} + \frac{1}{2}\boldsymbol{e}\right)\right],$$

where $\boldsymbol{v}_r = (\pi_1 q_{1r}, \dots, \pi_K q_{Kr})$. *Furthermore, the stationary expected cost is given by*

$$\mu := \lim_{t\to\infty} \mathbf{E}C_t = 1 + \sum_{k,r}^{*}\boldsymbol{v}_r(I-\hat{Q}_{rk})^{-1}Q_k\boldsymbol{e}. \tag{8}$$

Computing the stationary search cost by (8) involves inversions of $K \times K$ matrices instead of the $N \times N$ matrices required in the general Markov model. The number of inversions needed is normally large, so when K is sufficiently smaller than N, major reductions in computation time are possible.

Theorem 1 gives classical results as special cases. In particular, Markov-modulated requests become i.i.d. if we reduce the matrix P to the scalar 1 and put $q_{1r} = q_r := \mathbf{P}[R_t = r]$. Then $\hat{Q}_{rk} = 1 - q_r - q_k$, $Q_k = q_k$, and $\boldsymbol{v}_r = q_r$ are also scalars which when substituted into Theorem 1 give [2,11]

$$\mathbf{E}C_t = 1 + 2\sum_{r<k}\frac{q_r q_k}{q_r + q_k} + \sum_{r<k}\frac{(q_r - q_k)^2(1-q_r-q_k)^t}{2(q_r+q_k)}, \tag{9}$$

with the stationary mean

$$\mu = 1 + 2\sum_{r<k}\frac{q_r q_k}{q_r+q_k}. \tag{10}$$

If requests are generated by an aperiodic, irreducible, finite-state Markov chain, R, then the stationary expected search cost is given by [10]

$$\mu = 1 + \sum_{r,k}^{*}\frac{1}{m_{rk} + m_{kr}}, \tag{11}$$

where m_{ij} is the expected first passage time from state i to state j in R. To verify this, put $q_{ir} = \mathbf{1}\ (i = r)$ and $J = R$, and hence $K = N$. We have $\boldsymbol{v}_r = (0, \dots, \pi_r, \dots, 0)$ and $\hat{Q}_{rk} = \tilde{P}_{rk}$, where $\tilde{P}_{rk}$ is obtained from $\tilde{P}$ by replacing the rth and kth columns by zero columns. Similarly, $Q_k\boldsymbol{e}$ is a column vector with elements $\tilde{p}_{1k}, \dots, \tilde{p}_{Nk}$, so for each $r \neq k$, we have $\boldsymbol{v}_r(I-\hat{Q}_{rk})^{-1}Q_k\boldsymbol{e}$ as the probability of starting in state r and reaching state k before again visiting state r in the Markov chain $\tilde{P}$. Then, by Lemmas 3.1.1 and 3.1.2 of [10], it follows that $\boldsymbol{v}_r(I-\hat{Q}_{rk})^{-1}Q_k\boldsymbol{e} = 1/(m_{rk} + m_{kr})$, which when substituted into Theorem 1 gives (11).

To compute the second moment of C_t, we square (1) and obtain

$$C_t^2 = 1 + 3\sum_{i=1}^{K}\sum_{k,r}^{*}\mathbf{1}\ (J_t = i, R_t = r, A_t^{kr}) + \sum_{i=1}^{K}\sum_{k_1,k_2,r}^{*}1\ (J_t = i, R_t = r, A_t^{k_1 r}A_t^{k_2 r}). \tag{12}$$

Let $A_t^{k_1k_2r}$ be the event that the relative (left-to-right) ordering of k_1, k_2, and r in σ_t is k_1k_2r. We observe that $A_t^{k_1r}A_t^{k_2r}$ is the union of the two disjoint events $A_t^{k_1k_2r}$ and $A_t^{k_2k_1r}$, and then take the expected value of (12) to obtain

$$\mathbf{E}C_t^2 = 3\mathbf{E}C_t - 2 + \sum_{i=1}^{K} {\sum_{k_1,k_2,r}}^{*} \mathbf{P}[J_t = i, R_t = r, A_t^{k_1k_2r}] + \mathbf{P}[J_t = i, R_t = r, A_t^{k_2k_1r}]. \tag{13}$$

The probabilities $\mathbf{P}[J_t = i, R_t = r, A_t^{k_1k_2r}]$ in the expression above are computed by

$$\begin{aligned}
&\mathbf{P}[J_t = i, R_t = r, A_t^{k_1k_2r}]\\
&\quad = \sum_{m_1=1}^{t-1} \sum_{m_2=m_1+1}^{t} \mathbf{P}[J_t = i, R_t = r, R_{t-1} \notin \{k_1, k_2, r\}, \ldots, R_{t-m_1+1} \notin \{k_1, k_2, r\}, R_{t-m_1} = k_1,\\
&\qquad\qquad R_{t-m_1-1} \notin \{k_2, r\}, \ldots, R_{t-m_2+1} \notin \{k_2, r\}, R_{t-m_2} = k_2]\\
&\qquad + \sum_{m_1=1}^{t} \mathbf{P}[J_t = i, R_t = r, R_{t-1} \notin \{k_1, k_2, r\}, \ldots, R_{t-m_1+1} \notin \{k_1, k_2, r\}, R_{t-m_1} = k_1,\\
&\qquad\qquad R_{t-m_1-1} \notin \{k_2, r\}, \ldots, R_0 \notin \{k_2, r\}, A_0^{k_2,r}]\\
&\qquad + \mathbf{P}[J_t = i, R_t = r, R_s \notin \{k_1, k_2, r\}, 0 \leqslant s \leqslant t-1, A_0^{k_1k_2r}],
\end{aligned} \tag{14}$$

so we have

Theorem 2. *The stationary variance is given by*

$$\lim_{t\to\infty} \mathrm{Var}(C_t) = \mu - 1 - (\mu - 1)^2 + {\sum_{k_1,k_2,r}}^{*} \boldsymbol{v}_r (I - \hat{Q}_{rk_1k_2})^{-1} (Q_{k_1} (I - \hat{Q}_{rk_2})^{-1} Q_{k_2} + Q_{k_2} (I - \hat{Q}_{rk_1})^{-1} Q_{k_1}) \boldsymbol{e}. \tag{15}$$

Proof. From (14), a time-reversal argument, and (5), it follows that

$$\begin{aligned}
\lim_{t\to\infty} \mathbf{P}[R_t = r, A_t^{k_1k_2r}] &= \sum_{s_1=0}^{\infty} \sum_{s_2=0}^{\infty} \boldsymbol{v}_r \hat{Q}_{rk_1k_2}^{s_1} Q_{k_1} \hat{Q}_{rk_2}^{s_2} Q_{k_2} \boldsymbol{e}\\
&= \boldsymbol{v}_r (I - \hat{Q}_{rk_1k_2})^{-1} Q_{k_1} (I - \hat{Q}_{rk_2})^{-1} Q_{k_2} \boldsymbol{e}.
\end{aligned}$$

Substituting the expression above into (13) and then subtracting the square of the mean (μ^2) gives the result of the theorem. □

If requests are i.i.d. with $\mathbf{P}[R_t = r] = q_r$, then by specializing quantities as before, we get (recall that $q_r = \mathbf{P}[R_t = r]$)

$$\lim_{t\to\infty} \mathrm{Var}(C_t) = \mu - 1 - (\mu - 1)^2 + {\sum_{k_1,k_2,r}}^{*} \frac{q_r q_{k_1} q_{k_2} [(q_r + q_{k_1}) + (q_r + q_{k_2})]}{(q_r + q_{k_1})(q_r + q_{k_2})(q_r + q_{k_1} + q_{k_2})}. \tag{16}$$

For the general Markov model, the expression for the variance of the search cost was first derived in [13, Theorem 4.1]. That result can also be derived directly from Theorem 2, but we refrain from doing so, as the computation is quite awkward.

For a useful special case, assume that the list is partitioned into K disjoint subsets L_i, $l \leqslant i \leqslant K$, $L_i = \{(i,1), \ldots, (i, N_i)\}$, $\sum_{i=1}^{K} N_i = N$, and that, when the underlying Markov chain is in mode i ($J_t = i$), the request sequence can only access items from the subset L_i, i.e., $q_{i,(i,r)} > 0$ for $1 \leqslant r \leqslant N_i$ and $q_{j,(i,r)} = 0$ for $j \neq i$. With a small abuse of notation we write $q_{(i,r)} = q_{i,(i,r)}$.

Theorem 3. *For the stationary expected cost, we have*

$$\mu = 1 + \sum_i \pi_i \sum_{1 \leqslant k, l \leqslant N_i}^{*} \frac{q_{(i,k)} q_{(i,l)}}{q_{(i,k)} + q_{(i,l)}}$$

$$+ \sum_{i,j}^{*} \sum_{1 \leqslant k \leqslant N_i, 1 \leqslant l \leqslant N_j} \frac{\pi_i \pi_j q_{(i,k)} q_{(j,l)}}{\pi_i q_{(i,k)} + \pi_j q_{(j,l)} - \pi_i \pi_j q_{(i,k)} q_{(j,l)} (1/\pi_i + 1/\pi_j - m_{ij} - m_{ji})}, \tag{17}$$

where m_{ij} is the expected first passage time from state i to state j in J.

Remarks. Note that when the underlying Markov chain J represents an i.i.d. sequence (i.e., each row in P is equal to the distribution π), then $1/\pi_i + 1/\pi_j - m_{ij} - m_{ji} = 0$ and (17) reduces to (10). At the other extreme, if each subset reduces to $L_i = \{(i,1)\}$ and hence $J = R$, then (17) yields the result for Markovian requests stated in (11).

For the proof of Theorem 3, we first assemble a couple of well-known results for finite Markov chains. Let m_{ij} be the expected first passage time from state i to state j in J, and let ${}_j p_{ii}^{(n)}$ be the probability of going from state i to state i in n steps without visiting j (these are called taboo probabilities [8, p. 45]); let $\tilde{m}_{ij}$, ${}_j \tilde{p}_{ii}^{(n)}$ represent the analogous quantities for the reversed chain $\tilde{J}$. We denote by $\tilde{P}^{*i}$ the matrix obtained from $\tilde{P}$ by replacing its ith column by zero column; similarly, $\tilde{P}^{*ij}$ denotes the matrix resulting from the replacement of columns i and j by zeros.

Lemma 1. *Let $i \neq j$. Then*
(i) $\tilde{m}_{ii}/(\tilde{m}_{ij} + \tilde{m}_{ji}) = 1/(1 + \sum_{n=1}^{\infty} {}_j \tilde{p}_{ii}^{(n)}) = \det(I - \tilde{P}^{*i})/\det(I - \tilde{P}^{*ij})$,
(ii) $m_{jj} = \tilde{m}_{jj} = 1/\pi_j$,
(iii) $m_{ij} + m_{ji} = \tilde{m}_{ij} + \tilde{m}_{ji}$.

Proof. The first equality in (i) follows from [8, Eq. (14), p. 49, Corollary 1, p. 65]. The second equality is just an algebraic identity. Statements (ii) and (iii) represent Lemma 3.1.2 in [10]. □

Proof of Theorem 3. We will show that the expression in (17) follows directly from (8). First, note that $\boldsymbol{v}_{(i,k)} = (0, \ldots, 0, \pi_i q_{(i,k)}, 0, \ldots, 0)$, $(i,k) \in L_i$ and that $Q_{(i,k)}$ has only 1 nonzero column; it is the ith column of $\tilde{P}$ multiplied by $q_{(i,k)}$. Also, for $k \neq l$, we have $Q_{(i,k)(i,l)} = Q_{(i,k)} + Q_{(i,l)}$ and $\hat{Q}_{(i,k)(i,l)} = P - Q_{(i,k)(i,l)}$. These observations and simple algebra yield

$$\boldsymbol{v}_{(i,k)} (I - \hat{Q}_{(i,k)(i,l)})^{-1} Q_{(i,l)} \boldsymbol{e} = \pi_i \frac{q_{(i,k)} q_{(i,l)}}{q_{(i,k)} + q_{(i,l)}},$$

this justifies the first sum in (17).

When a pair of items (i,k) and (j,l) belong to different subsets ($i \neq j$), then $q_{(i,k)(j,l)}$ has two nonzero columns which are the ith and jth columns in P multiplied by $q_{(i,k)}$ amd $q_{(j,l)}$, respectively; also, we have $\hat{Q}_{(i,k)(j,l)} = P - Q_{(i,k)(j,l)}$. Then

$$\boldsymbol{v}_{(i,k)} (I - \hat{Q}_{(i,k)(j,l)})^{-1} Q_{(j,l)} \boldsymbol{e} = \frac{\pi_i q_{(i,k)} q_{(j,l)}}{\det(I - \hat{Q}_{(i,k)(j,l)})} \sum_{j_1=1}^{K} (-1)^{i+j_1} \tilde{p}_{j_1 j} \det((I - \hat{Q}_{(i,k)(j,l)})_{*j_1}^{*i})$$

$$= \pi_i q_{(i,k)} q_{(j,l)} \frac{\det(I - \tilde{P}^{*j})}{\det(I - \hat{Q}_{(i,k)(j,l)})}, \tag{18}$$

where $(I-\hat{Q}_{(i,k)(j,l)})^{*i}_{*j_1}$ is the matrix obtained from $(I-\hat{Q}_{(i,k)(j,l)})$ by deleting its j_1th row and ith column. The first equality in (18) just exploits the determinant representation of $(I-\hat{Q}_{(i,k)(j,l)})^{-1}$; the second equality follows from elementary properties of determinants, which also give us

$$\det(I-\hat{Q}_{(i,k)(j,l)}) = q_{(i,k)}\det(I-\tilde{P}^{*i}) + q_{(j,l)}\det(I-\tilde{P}^{*j}) \\ -q_{(i,k)}q_{(j,l)}(\det(I-\tilde{P}^{*i}) + \det(I-\tilde{P}^{*j}) - \det(I-\tilde{P}^{*ij})). \tag{19}$$

Next, let D_P be a determinant obtained by replacing one column in $(I-\tilde{P})$ with $\boldsymbol{e}$ (the value of D_P is independent of the column replaced). It is easy to show that

$$\pi_i = \frac{\det(I-\tilde{P}^{*i})}{D_P}. \tag{20}$$

Thus, after dividing (19) by D_P, and applying Lemma 1 and (20), we arrive at

$$\frac{\det(I-\hat{Q}_{(i,k)(j,l)})}{D_P} = q_{(i,k)}\pi_i + q_{(j,l)}\pi_j - q_{(i,k)}q_{(j,l)}(\pi_i+\pi_j-\pi_i\pi_j(m_{ij}+m_{ji})). \tag{21}$$

Finally, divide the numerator and denominator in (18) by D_P and substitute (20) and (21) to obtain

$$\boldsymbol{v}_{(i,k)}(I-\hat{Q}_{(i,k)(j,l)})^{-1}Q_{(j,l)}\boldsymbol{e} = \frac{\pi_i\pi_j q_{(i,k)}q_{(j,l)}}{\pi_i q_{(i,k)} + \pi_j q_{(j,l)} - \pi_i\pi_j q_{(i,k)}q_{(j,l)}(1/\pi_i + 1/\pi_j - m_{ij} - m_{ji})}$$

which, when summed over i,j,k,l yields the second sum in (17). □

3. Examples and discussion

As we have noted, in many applications, a state of the modulating process J determines that mode or context, of perhaps many, in which a list is being accessed. As a symmetric example for $K=2$ possible accessing modes, consider a modulated Zipf's law. In mode 1, the request frequencies are $q_{1r}=1/(rH_N)$, $1\leqslant r\leqslant N$, with $H_N=\sum_{i=1}^{N} 1/r$. For maximum contrast in mode 2, we reverse the ordering of L by request frequency, i.e., we take the complementary probabilities $q_{2r}=1/(N-r+1)\times 1/H_N$, $1\leqslant r\leqslant N$. The two-state Markov chain J is defined by the transition probabilities $p_{12}=p_{21}=(1-w)/2$, where w is a 'memory' parameter with $|w|\leqslant 1$. As w decreases to -1, R increases its tendency to jump from one mode to the other, whereas it resides for long periods in a mode when w is close to 1. The stationary distribution of J is given by $\pi_1=\pi_2=1/2$, and the unconditional request probabilities are $q_r=\pi_1 q_{1r}+\pi_2 q_{2r}=(q_{1r}+q_{2r})/2$. Note that $w=0$ is the i.i.d. case with probabilities q_r.

Fig. 1 plots the expected search cost (8) for $N=50,100$ as a function of w, and compares it to two i.i.d. models. The dashed line labeled 'MTF i.i.d.' refers to the MTF performance under independent requests each drawn from $\{q_r\}$. If the request frequencies are known in advance, and if requests are independent, then it is optimal to order L by decreasing request frequency and to keep this ordering fixed. This gives the 'optimal static i.i.d.' dashed line in the figure. As can be seen, for negative memory (w), MTF performs nearly as it would were requests i.i.d. . And for positive memory, the MTF expected search cost experiences a steep drop, especially as w nears 1. Indeed, for w sufficiently close to 1, MTF does even better than in the optimal static i.i.d. case. In this regime, MTF normally spends very long periods of time in making i.i.d. requests according to $\{q_{1r}\}$ or $\{q_{2r}\}$ before switching from one mode to the other; relatively little time is spent in making the major list restructuring that accompanies changes in mode. And as implied by the figure, when w is close to 1, MTF operating in either mode is better than the optimal static algorithm in the single 'combined' mode with independent requests drawn form $\{q_r\}$, $q_r=(q_{1r}+q_{2r})/2$.

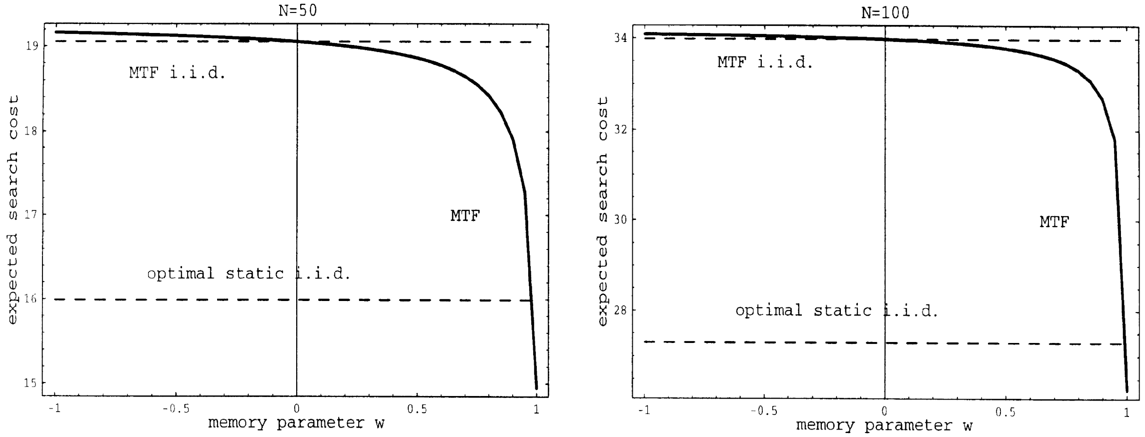

Fig. 1. Illustration for the Markov-modulated Zipf’s law.

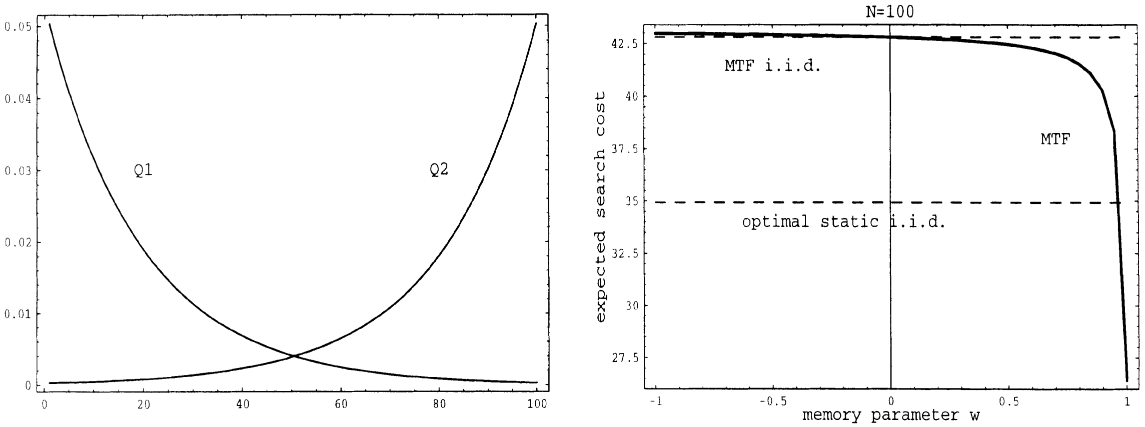

Fig. 2. Illustration for Markov-modulated geometric request sequences.

Next, consider a similar experiment with Markov modulated geometric request sequences. The probabilities are

$$q_{1r} = \lambda^{(r-1)}/c_N, \qquad q_{2r} = \lambda^{(N-r)}/c_N, \quad 1 \leqslant r \leqslant N,$$

with the normalization constant $c_N = \sum_{r=0}^{N-1} \lambda^r$.

For $N = 100$, and $\lambda = 0.95$ the two distributions $Q1$ and $Q2$ are pictured on the left-hand side of Fig. 2. On the right-hand side, we have displayed the dependence of the expected search cost on the memory parameter w. It is interesting to observe once again that for w close to one both ‘MTF i.i.d.’ and ‘optimal static i.i.d.’ are quite pessimistic in comparison with the values of the MTF algorithm in which locality of reference is modeled explicitly.

All of the observations made to this point apply also to our third and final example in which we assume that the list is partitioned into two equal-size blocks; Markov modulated Poisson request sequences are drawn alternately from the two blocks. Specifically, we choose $K = 2$, $N_1 = N_2 = N/2 = 20$, and take

$$q_{(1,k)} = q_{(2,k)} = \lambda^{(k-1)}/(c_N(k-1)!), \quad 1 \leqslant k \leqslant N/2,$$

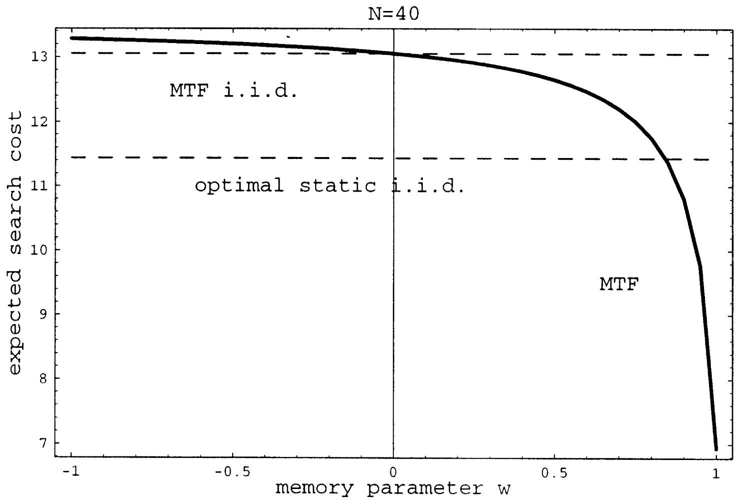

Fig. 3. Illustration for Markov-modulated Poisson request sequences.

with the normalization constant $c_N = \sum_{r=0}^{N/2-1} \lambda^k/k!$. Here, we use formula (17) with

$$m_{12} + m_{21} = \frac{1}{p_{12}} + \frac{1}{p_{21}} = \frac{4}{1-w}$$

to compute the expected search cost. For $\lambda = 10$, this computation is shown in Fig. 3.

4. Final remarks

The stationary distribution of search cost is of obvious interest, especially in studies of LRU caching where tail probabilities (fault rates) are needed. To obtain a formal solution to this problem, assume that both the MMR process and the search-cost process $\{C_t\}$ are stationary (initial states are samples from the stationary distributions). The derivation below follows closely that in [3], so we will be brief. For $k \geqslant 1$, let $X_k = \{\tilde{R}_1, \ldots, \tilde{R}_k\}$ be the set of distinct elements in the first k requests and $|X_k|$ be the cardinality of X_k. Then

$$\begin{aligned}\mathbf{P}[R_0 = r, C_0 = n] &= \sum_{k=0}^{\infty} \mathbf{P}[\tilde{R}_0 = r, r \notin X_r, |X_k| = n-1, \tilde{R}_{k+1} = r] \\ &= \sum_{k=0}^{\infty} \sum_{\{B:|B|=n-1, r\notin B\}} \mathbf{P}[\tilde{R}_0 = r, X_k = B, \tilde{R}_{k+1} = r]. \end{aligned} \tag{22}$$

By the inclusion–exclusion formula,

$$\begin{aligned}\mathbf{P}[\tilde{R}_0 = r, X_k = B, \tilde{R}_{k+1} = r] &= \sum_{\{A:A\subseteq B\}} (-1)^{|B-A|} \mathbf{P}[\tilde{R}_0 = r, X_k \subseteq A, \tilde{R}_{k+1} = r] \\ &= \sum_{\{A:A\subseteq B\}} (-1)^{|B-A|} \boldsymbol{v}_r Q^k(A) Q_r \boldsymbol{e}. \end{aligned} \tag{23}$$

Substituting (23) into (22) gives

$$\mathbf{P}[R_0 = r, C_0 = n] = \sum_{\{B:|B|=n-1,r\notin B\}} \sum_{\{A\,:\,A\subseteq B\}} (-1)^{|B-A|}\boldsymbol{v}_r(I-Q(A))^{-1}Q_r\boldsymbol{e}$$

$$= \sum_{\{A:|A|\leqslant n-1,r\notin A\}} \sum_{\{B:B\supseteq A,|B|=n-1,r\notin B\}} (-1)^{|B-A|}\boldsymbol{v}_r(I-Q(A))^{-1}Q_r\boldsymbol{e}.$$

The number of sets in $\{B\colon B\supseteq A, |B|=n-1, |A|=a, r\notin B\}$ is $\binom{N-1-a}{n-1-a}$, so after simplifying and summing over r, we obtain the following result, which reduces easily to the result in [3] when requests are i.i.d.

Theorem 4.

$$\mathbf{P}[C_0=n] = \sum_{r=1}^{N}\sum_{a=0}^{n-1}(-1)^{n-a-1}\binom{N-1-a}{n-1-a}\sum_{\{A:|A|=a,r\notin A\}}\boldsymbol{v}_r(I-Q(A))^{-1}Q_r\boldsymbol{e}.$$

Unfortunately, Theorem 4 does not give us a feasible computation for interesting values of N (e.g., for N in the thousands, at least). We leave as an interesting open question the problem of estimating search-cost probabilities.

References

[1] J.L. Bentley, C.C. McGeoch, Amortized analysis of self-organizing sequential search heuristics, Commun. ACM 28 (4) (1985) 404–411.

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

[3] P.J. Burville, J.F.C. Kingman, On a model for storage and search, J. Appl. Probab. 10 (1973) 697–701.

[4] E.G. Coffman Jr., P.J. Denning, Operating Systems Theory, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[5] R.P. Dobrow, J.A. Fill, The move-to-front rule for self-organizing lists with Markov dependent requests, in: D. Aldous, P. Diaconis, J. Spencer, J.M. Steele (Eds.), Discrete Probability and Algorithms, Springer, Berlin, 1995, pp. 57–80.

[6] B.M. Hochwald, P.R. Jelenković, State learning and mixing in hidden Markov models and the Gilbert–Elliot channel, IEEE Trans. Inform. Theory 45 (1) (1999) 128–138.

[7] P.R. Jelenković, Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities, Ann. Appl. Probab. 9 (2) (1999) 430–464.

[8] Kai Lai Chung, Markov Chains with Stationary Transition Probabilities, Springer, New York, 1960.

[9] D.E. Knuth, The Art of Computer Programming, Vol. 3: Sorting and Searching, Addison-Wesley, Reading, MA, 1973.

[10] K. Lam, M.Y. Leung, M.K. Siu, Self-organizing files with dependent accesses, J. Appl. Probab. 21 (1984) 343–359.

[11] J. McCabe, On serial files with relocatable records, Oper. Res. 13 (1965) 609–618.

[12] E.R. Rodrigues, Convergence to stationary state for a Markov move-to-front scheme, J. Appl. Probab. 32 (3) (1995) 768–776.

[13] E.R. Rodrigues, The performance of the move-to-front scheme under some particular forms of Markov requests, J. Appl. Probab. 32 (4) (1995) 1089–1102.

[14] E. Seneta, Non-negative Matrices and Markov Chains, Springer, Berlin, 1981.