Least-Recently-Used Caching with Dependent Requests

Predrag Jelenković and Ana Radovanović
Department of Electrical Engineering
Columbia University
New York, NY 10027, USA
{predrag, anar}@ee.columbia.edu
tel.: (1 212) 854 8174, 939 7157
fax: + (1 212) 932 9421

August 2002

Abstract

We investigate a widely popular Least-Recently-Used (LRU) cache replacement algorithm with semi-Markov modulated requests. Semi-Markov processes provide the flexibility for modeling strong statistical correlation, including the broadly reported long-range dependence in the World Wide Web page request patterns. When the frequency of requesting a page n is equal to the generalized Zipf's law $c/n^{\alpha}, \alpha > 1$, our main result shows that the cache fault probability is asymptotically, for large cache sizes, the same as in the corresponding LRU system with i.i.d. requests. This appears to be the first explicit average case analysis of LRU caching with statistically dependent request sequences. The surprising insensitivity of LRU caching performance demonstrates its robustness to changes in document popularity. Furthermore, we show that the derived asymptotic result and simulation experiments are in excellent agreement, even for relatively small cache sizes.

Keywords: least-recently-used caching, Zipf's law, semi-Markov processes, heavy-tailed distributions, long-range dependence, average-case analysis

This work is supported by the NSF Grant No. 0092113.

1 Introduction

The basic idea of caching is to maintain high-speed access to a subset of k items out of a larger collection of N documents that cannot be accessed quickly. Originally, caching was used in computer systems to speed-up the data transfer between the central processor unit and slow local memory. The renewed interest in caching stems from its application to increasing the speed of accessing Internet Web documents.

One of the fundamental issues of caching is the problem of selecting and possibly dynamically updating the k items that need to be stored in the fast memory (cache). The optimal solution to this problem is often very difficult to find and, therefore, a number of heuristic, usually dynamic, cache updating algorithms have been proposed. Among the most popular algorithms are those based on the Least-Recently-Used (LRU) cache replacement rule. The wide popularity of this rule is primarily due to its high performance and ease of implementation. LRU algorithm tends to both keep more frequent items in the cache as well as quickly adapt to the potential changes in document popularity, resulting in efficient performance.

In order to further the insight into designing network caching algorithms, it is important to gain a thorough understanding of the baseline LRU cache replacement policy. Basic references on the performance analysis of caching algorithms can be found in Section 6 of Knuth [17]. In the analysis of LRU caching scheme there have been two approaches: combinatorial and probabilistic studies. For the combinatorial (amortized, competitive) analysis the reader is referred to Bentley and McGeoch [3] and Sleator and Tarjan [23]. Recent results and references for this approach can be found in Borodin et al. [5] and Irani et al. [13]. In this paper we focus on the average-case or probabilistic analysis.

Early work on the probabilistic analysis of LRU caching, and the related Move-To-Front (MTF) searching, algorithm with i.i.d. requests dates back to McCabe [18]. This work has been followed by investigations of Burville and Kingman [6], Rivest [21], Bitner[4], Phatarford [20], Fill [11], Flajolet et al. [12] and others; a more extensive list of references and brief historical overview can be found in [15].

Recently, for the independent reference model, in [15] a new analytically tractable asymptotic approximation technique of the LRU fault probability was developed. However, an equivalent understanding of LRU performance with statistically dependent request sequences is still lacking. Several papers, including Rodrigues [22], Dobrow and Fill [9] and Coffman and Jelenković [8], develop representation results for the LRU cache fault probability, but these results appear to be computationally intractable, as pointed out in [8]. Despite the lack of analytical tractability, numerous empirical studies, e.g. see [1], emphasize the importance of understanding the caching behavior in the presence of strong statistical correlation, including the long-range dependence.

In order to alleviate the preceding problem, this paper provides the first explicit asymptotic characterization of the LRU cache fault probability in the case of statistically dependent requests. Our doubly stochastic Poisson reference model, capable of capturing a broad range of statistical correlation, is described in the following section. Using this model and the Poisson decomposition/superposition properties, similarly as in Fill [10], in Section 3 we develop a representation theorem for the stationary search cost distribution. This representation theorem provides a starting point for our new large deviation analysis that, for the case of generalized Zipf's law requests, yields our main results stated in Theorems 2 and 3.

Informally, our main results show that the LRU fault probability is asymptotically invariant to the underlying dependency structure of the modulating process, i.e., for large cache sizes, the LRU fault probability behaves exactly the same as in the case of independent request sequences [15]. This may appear surprising given the impact that the statistical correlation has on the asymptotic performance of queuing models. Furthermore, in Section 5 extensive numerical experiments show an excellent agreement between our analytical results and simulations. The paper is concluded in Section 6 with a brief discussion on the impact of our findings on designing network caching systems.

2 Model description

Consider N items, out of which k are kept in a fast memory (cache) and the remaining $N-k$ are stored in a slow memory. Each time a request for an item is made, the cache is searched first. If the item is not found there, it is brought in from the slow memory and replaced with the least recently accessed item from the cache. Such a replacement policy is commonly referred to as LRU, as previously stated in the introduction. The performance quantity of interest for this algorithm is the LRU fault probability, i.e. the probability that the requested item is not in the cache. Our goal in this paper is to asymptotically characterize this probability.

The fault probability of the LRU caching is equivalent to the tail of the searching cost distribution for the MTF searching algorithm. In order to justify this claim, we note that k elements in the cache, under the LRU rule, are arranged in the increasing order of their last access times. Each time there is a request for an item that is not in the cache, the item is brought to the first position of the cache and the last element of the cache is moved to the slow memory. We argue that the fault probability stays the same if the remaining $N-k$ items in the slow memory are arranged in any specific order. In particular, they can be arranged in the increasing order of their last access times. The obtained algorithm is then the same as the MTF searching algorithm. Additional arguments that justify the connection between the MTF search cost distribution and LRU cache fault probability can be found in [12], [10] and [15]. Hence, we proceed with a description of the MTF algorithm.

More formally, consider a finite list of items $L = \{1, \dots, N\}$, and a sequence of requests that arrive according to a sequence of Poisson points $\{\tau_n, -\infty < n < \infty\}$ of unit rate. At each point τ_n, we use R_n to denote a document that has been requested, i.e. an event $\{R_n = i\}$ represents a request for document i; we assume that the sequence $\{R_n\}$ is independent of the arrival Poisson points $\{\tau_n\}$. The dynamics of the MTF algorithm is defined as follows. Suppose that the system starts at moment τ_0 of 0th request with an initial permutation of the list Π_0. Then, at every time instant τ_n, $n \geq 0$, that an item, say i, is requested, its position in the list is first determined; if i is in the kth position we say that the search cost C_n^N for this item is equal to k. Now, the list is updated by moving item i to the first position of the list and items in positions $1, \dots, k-1$, are moved one position down. Note that, according to the discussion in the preceding paragraph, $\mathbb{P}[C^N > k]$ represents the stationary fault probability for a cache of size k.

In the remaining part of this section we describe the dependency structure of the request sequence $\{R_n\}$. Let $\{T_n, -\infty < n < \infty\}$, $T_0 \leq 0 < T_1$, be a point process with almost surely (a.s.) strictly increasing points $(T_{n+1} > T_n)$ and $\{J_{T_n}, -\infty < n < \infty\}$ a finite state space process taking values in $\{1, \dots, M\}$. Then, we construct a piecewise constant right-continuous *modulating process* J_t, as

$$J_t = J_{T_n}, \quad \text{if} \qquad T_n \leq t < T_{n+1}.$$

We assume that J_t is stationary and ergodic with stationary distribution $\pi_k = \mathbb{P}[J_t = k]$ and independent of Poisson points $\{\tau_n\}$. Next, for any $k, m \leq M$, due to ergodicity

$$\mathbb{P}[J_t = k | J_0 = m] \to \pi_k \qquad \text{as} \qquad t \to \infty. \tag{1}$$

To avoid trivialities, we assume that $\min_k \pi_k > 0$.

For each $1 \leq k \leq M$, let $q_i^{(k)}, i \geq 1$ be a probability mass function. Next, the dynamics of R_n is uniquely determined by the modulating process J_t according to the following equation

$$\mathbb{P}[R_l = i_l, 1 \leq l \leq n | J_t, t \leq \tau_n] = \prod_{l=1}^{n} q_{i_l}^{(J_{\tau_l})} \qquad n \geq 1; \tag{2}$$

i. e., the sequence of requests R_n is conditionally independent given the modulating process J_t. Therefore, the constructed request process $\{R_n\}$ is stationary and ergodic as well. We will use

$$q_i = \mathbb{P}[R = i] = \sum_{k=1}^{M} \pi_k q_i^{(k)}$$

to express the marginal request distribution with the assumption that $q_i > 0$ for all $i \geq 1$. The preceding processes are constructed on a probability space $(\Omega, \sigma(\Omega), \mathbb{P})$.

3 Preliminary results

In this section we first provide, in Lemma 1, general conditions under which the search cost process C_n^N converges to stationarity. Then, in the following subsection we characterize the stationary search cost distribution in Theorem 1 and Proposition 1. The remaining part of the section contains the results on MTF searching with i.i.d. requests that will be used in proving our main theorems.

Lemma 1 *If the request process R_n is stationary and ergodic, then for any initial permutation Π_0 of the list, the search cost process C_n^N converges in distribution to C^N as $n \to \infty$, where*

$$C^N \triangleq \sum_{i=1}^{N} \sum_{m=1}^{\infty} (1 + S_i(m-1)) 1[R_{-m} = i, \mathcal{R}_i(m-1), R_0 = i],$$

$S_i(m)$ is the number of distinct items, different from i, among $R_{-m}, \ldots, R_{-1}$ and event $\mathcal{R}_i(m) \triangleq \{R_{-j} \neq i, 1 \leq j \leq m\}$, $m \geq 1$; $S(0) \equiv 0$, $\mathcal{R}_i(0) \equiv \Omega$.

Proof: For simplicity let $C_n \equiv C_n^N$. Note that, due to the stationarity of the request process $\{R_n\}$, C_n is equal in distribution to the search cost $C_0^{(n)}$ at the moment of 0th request τ_0, given that the MTF process started at time τ_{-n} with initial permutation Π_0. Now, each of the summands of the following identity

$$C_0^{(n)} = \sum_{i=1}^{N} C_0^{(n)} 1[R_0 = i] \tag{3}$$

can be represented as

$$C_0^{(n)} 1[R_0 = i] = \sum_{m=1}^{n} (1 + S_i(m-1)) 1[R_{-m} = i, \mathcal{R}_i(m-1), R_0 = i] + C_0^{(n)} 1[\mathcal{R}_i(n), R_0 = i], \tag{4}$$

since $C_0^{(n)} = 1 + S_i(m-1)$ on event $\{R_{-m} = i, \mathcal{R}_i(m-1), R_0 = i\}$. The second term in the preceding equality is bounded by $N1[\mathcal{R}_i(n)]$, which, by ergodicity, satisfies a.s.

$$\lim_{n \to \infty} N 1[\mathcal{R}_i(n)] = 0.$$

Thus, the last limit, monotonicity of the sum in (4) and identity (3) imply that $C_0^{(n)}$ converges a.s. to C^N as $n \to \infty$. Therefore, C_n^N converges in distribution to C^N as $n \to \infty$. $\diamond$

3.1 Representation theorem

At this point, we will derive a representation theorem for the stationary search cost C^N, as defined in Lemma 1. Note that C^N is uniquely defined by the search cost process $\{R_n, n \leq 0\}$ and, therefore, it implicitly depends on $\{J_{\tau_0+t}, t \leq 0\}$. However, since τ_0 is independent from $\{J_t\}$, the process $\{J_{\tau_0+t}, t \leq 0\}$ is equal in distribution to $\{J_t, t \leq 0\}$. Thus, without loss of generality we can set $\tau_0 = 0$. Next, let τ^i_{-1} be the last moment of time $t < 0$ that item i was requested. Then, an equivalent continuous time representation of C^N is equal to

$$C^N = \sum_{i=1}^{N} (1 + S_i(\tau^i_{-1}; J)) 1[R_0 = i],$$

where, similarly as in Lemma 1, $S_i(t; J)$ represents the number of distinct items, different than i, that are requested in interval $[-t, 0)$. Now, using double conditioning and the last identity, we arrive at

$$\mathbb{P}[C^N > x] = \mathbb{E} \int_0^\infty \sum_{i=1}^{N} \mathbb{P}_{\sigma_t} \left[S_i(t; J) > x - 1, R_0 = i, \tau^i_{-1} \in (-t, -t + dt) \right],$$

where σ_t is the σ-algebra $\sigma(J_u, -t \leq u \leq 0)$ and $\mathbb{P}_{\sigma_t}[\cdot] = \mathbb{P}[\cdot|\sigma_t]$. Using the fact that the request process R_n, by (2), is conditionally independent given the modulating process J_t and that the variables $S_i(t; J)$ and τ^i_{-1} are uniquely determined by the values of $\{R_n, n \leq -1\}$ and the Poisson arrivals for $t < 0$ we conclude that R_0 is conditionally independent from $S_i(t; J)$ and τ^i_{-1}, given σ_t, and thus

$$\mathbb{P}[C^N > x] = \mathbb{E} \int_0^\infty \sum_{i=1}^{N} q_i^{(J_0)} \mathbb{P}_{\sigma_t} \left[S_i(t; J) > x - 1, \tau^i_{-1} \in (-t, -t + dt) \right]. \tag{5}$$

Next, we intend to show that variables $S_i(t; J)$ and τ^i_{-1} are conditionally independent given σ_t. To this end, we exploit the Poisson superposition/decomposition properties of the arrival process. Let $N_j(u; J)$ be the number of requests for item j in $[-u, 0), 0 < u \leq t$ and $B_j(t; J) = 1[N_j(t; J) > 0]$. Then, $S_i(t; J)$ can be represented as

$$S_i(t; J) = \sum_{j \neq i, 1 \leq j \leq N} B_j(t; J). \tag{6}$$

Now, we show that, for different j, processes $\{N_j(u; J), 0 < u \leq t\}$ are mutually independent Poisson processes given σ_t. In this regard, for any $t > u > 0$, let V_n be an interval in $[-u, 0)$ on which the modulating process stays constant, i.e.

$$V_n = T_{n+1} \wedge 0 - T_n \vee (-u),$$

where $a \wedge b \equiv \min(a, b)$ and $a \vee b \equiv \max(a, b)$. Since, by (2), the request process is conditionally independent given σ_t, and independent from the Poisson arrival points, the Poisson decomposition theorem (see Section 4.5 of [7]) implies that the number of requests for item j in an interval V_n, given σ_t, is a Poisson variable with expected value $q_j^{(J_{T_n \vee (-u)})} V_n$. Furthermore, the Poisson variables for different j and different intervals V_n are independent given σ_t. Thus, given σ_t, aggregating the independent Poisson requests for item j over all intervals $V_n \subset [-u, 0]$, by Poisson superposition theorem (see Section 4.4 of [7]), shows that $N_j(u; J)$ are mutually independent Poisson variables for different j. Furthermore, by repeating the preceding arguments over an arbitrary set of disjoint intervals $[-u_m, -u_{m-1}), \ldots, [-u_1, 0), 0 < u_1 \leq \cdots \leq u_{m-1} \leq u_m \leq t$, it easily follows that, for different j, $\{N_j(u; J), 0 < u \leq t\}$ are mutually

independent Poisson processes given σ_t. In particular, for any fixed t, the Bernoulli variables $B_j(t;J)$ are conditionally independent given σ_t with

$$\mathbb{P}_{\sigma_t}[B_j(t;J)=1] = 1 - e^{-\hat{q}_j t}, \tag{7}$$

where

$$\hat{q}_j = \sum_{k=1}^{M} q_j^{(k)} \hat{\pi}_k \text{ and } \hat{\pi}_k = \frac{1}{t}\int_{-t}^{0} 1[J_u = k]\, du. \tag{8}$$

Therefore, since $\{\tau_{-1}^i > t\} = \{N_i(t;J) = 0\}$, the conditional independence of variables $N_j(t;J)$, and equation (6) show that $S_i(t;J)$ and τ_{-1}^i are conditionally independent given σ_t. Using this fact and

$$\begin{aligned}\mathbb{P}_{\sigma_t}[\tau_{-1}^i \in (-t,-t+dt)] &= \mathbb{P}_{\sigma_t}[N_i(t-dt;J)=0, N_i(t;J) - N_i(t-dt;J) = 1] \\ &= e^{-\hat{q}_i t} q_i^{(J_{-t})} dt\end{aligned}$$

in (5) we derive the following representation theorem

Theorem 1 *The stationary distribution of the searching cost C^N satisfies*

$$\mathbb{P}[C^N > x] = \mathbb{E}\int_0^\infty \sum_{i=1}^{N} q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t} \mathbb{P}_{\sigma_t}[S_i(t;J) > x-1] dt, \tag{9}$$

with $S_j(t;J), B_j(t;J)$ and $\hat{q}_j$ satisfying equations (6), (7) and (8), respectively.

Remark 1 *Throughout this paper we will use the property that variables $S_j(t;J), B_j(t;J), j \geq 1$ are monotonically increasing in t and $B_j(t;J), j \geq 1$ are conditionally independent given σ_t.*

In the proposition that follows we investigate the limiting search cost distribution when the number of items $N \to \infty$. Now, assume that the probability mass functions $q_i^{(k)}, 1 \leq k \leq M$ are defined for all $i \geq 1$ with $\max q_i^{(k)} > 0$. Using these probabilities, for a given modulating process J_t and each $1 \leq N \leq \infty$ we define a sequence of request processes $\{R_n^N\}$, whose conditional request probabilities are equal to

$$q_{i,N}^{(k)} = \frac{q_i^{(k)}}{\sum_{i=1}^N q_i^{(k)}}, \quad 1 \leq i \leq N;$$

then, for each finite N, let C^N be the corresponding stationary search cost. In the case of the limiting request process $R_n = R_n^\infty$, similarly as in (6), introduce $S_i(t;J) = \sum_{j \neq i} B_j(t;J)$ to be equal to the number of different items, not equal to i, that are requested in $[-t,0)$; $B_j(t;J)$ is the Bernoulli variable representing the event that item j was requested at least once in $[-t,0)$. Now, we prove the limiting representation result that provides a starting point for our large deviation analysis in Section 4.

Proposition 1 *The constructed sequence of stationary search costs C^N converges in distribution to C as $N \to \infty$, where the distribution of C is given by*

$$\mathbb{P}[C > x] = \mathbb{E}\int_0^\infty \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t} \mathbb{P}_{\sigma_t}[S_i(t;J) > x-1]\, dt. \tag{10}$$

Remark 2 *For the i.i.d. case, this result was proved in Proposition 4.4 of [10].*

Proof: In order to prove the convergence in distribution, it is enough to show the pointwise convergence of distribution functions, i.e. for any $x \geq 0$, $\mathbb{P}[C^N > x] \to \mathbb{P}[C > x]$ as $N \to \infty$. This is easily achieved using the Dominated Convergence Theorem. For details see the appendix.

3.2 Results for i.i.d. requests

In this section we state several results that consider LRU caching scheme with independent requests that will be used in proving our main results. The MTF model with i.i.d. requests follows from our general problem formulation when the modulating process is assumed to be a constant, i.e. $J_t \equiv constant$. In this case the Bernoulli variables $\{B_j(t), j \geq 1\}$ that indicate that an item j was requested in $[-t, 0)$ are independent with success probabilities $\mathbb{P}[B_i(t) = 1] = 1 - e^{-q_i t}$. Then, using the notation $S_i(t) = \sum_{j \neq i} B_j(t)$, it is easy to see that the distribution of the limiting stationary search cost C from Proposition 1 reduces to

$$\mathbb{P}[C > x] = \int_0^\infty \sum_{i=1}^\infty q_i^2 e^{-q_i t} \mathbb{P}[S_i(t) > x - 1] dt. \tag{11}$$

The following two results, originally proved in Lemmas 1 and 2 of [14], are restated here for convenience. In this paper we are using the following standard notation. For any two real functions $a(t)$ and $b(t)$ and fixed $t_0 \in \mathbb{R} \cup \{\infty\}$ we will use $a(t) \sim b(t)$ as $t \to t_0$ to denote $\lim_{t \to t_0} a(t)/b(t) = 1$. Similarly, we say that $a(t) \gtrsim b(t)$ as $t \to t_0$, if $\liminf_{t \to t_0} a(t)/b(t) \geq 1$; $a(t) \lesssim b(t)$ has a complementary definition.

Lemma 2 *Assume that $q_i \sim c/i^\alpha$ as $i \to \infty$, with $\alpha > 1$ and $c > 0$. Then, as $t \to \infty$*

$$\sum_{i=1}^\infty (q_i)^2 e^{-q_i t} \sim \frac{c^{\frac{1}{\alpha}}}{\alpha} \Gamma\left(2 - \frac{1}{\alpha}\right) t^{-2+\frac{1}{\alpha}},$$

where Γ is the Gamma function.

Lemma 3 *Let $S(t) = \sum_{i=1}^\infty B_i(t)$ and $q_i \sim c/i^\alpha$ as $i \to \infty$, with $\alpha > 1$ and $c > 0$. Then, as $t \to \infty$*

$$m(t) \triangleq \mathbb{E}S(t) \sim \Gamma\left(1 - \frac{1}{\alpha}\right) c^{\frac{1}{\alpha}} t^{\frac{1}{\alpha}}.$$

Throughout the paper H denotes a sufficiently large positive constant, while h denotes a sufficiently small positive constant. The values of H and h are generally different in different places. For example, $H/2 = H$, $H^2 = H$, $H + 1 = H$, etc. The next straightforward lemma will be repeatedly used in the paper.

Lemma 4 *Let $\{B_i, i \geq 1\}$ be a sequence of independent Bernoulli random variables, $S = \sum_{i=1}^\infty B_i$ and $m = \mathbb{E}[S]$. Then for any $\epsilon > 0$, there exists $\theta_\epsilon > 0$, such that*

$$\mathbb{P}[|S - m| > m\epsilon] \leq H e^{-\theta_\epsilon m}.$$

The **proof** is given in the appendix. ◇

Now, we provide a general bound on the search cost distribution for the case when the request probabilities are polynomially bounded. In the following two lemmas, we also allow for some of the q_i-s to be equal to zero.

Lemma 5 *If $0 \leq q_i \leq H/i^\alpha$, then for any $x \geq 1$*

$$\mathbb{P}[C > x] \leq \frac{H}{x^{\alpha - 1}}.$$

Proof: If there were finitely many q_is that are positive, then we could always find a large enough cache size such that the fault probability is equal to zero and the bound trivially holds. Hence, without loss of generality we can assume that $q_i > 0$ for countably many $i \geq 1$. Therefore, $m(t) \nearrow \infty$ monotonicaly as $t \nearrow \infty$, implying that the inverse $m^{-1}(t)$ exists for any $t \geq 0$. Next, define $x_\epsilon = (1-\epsilon)(x-1)$, for some $0 < \epsilon < 1$. Now, using $S_i(t) \leq S(t)$ in (11), we derive

$$\begin{aligned}\mathbb{P}[C > x] &\leq \int_0^{m^{-1}(x_\epsilon)} \sum_{i=1}^{\infty} q_i^2 e^{-q_i t} \mathbb{P}[S(t) > x-1]\, dt + \int_{m^{-1}(x_\epsilon)}^{\infty} \sum_{i=1}^{\infty} q_i^2 e^{-q_i t}\, dt \\ &\triangleq I_1(x) + I_2(x).\end{aligned}$$

Then, since $S(t)$ is non-decreasing function in t

$$\begin{aligned}I_1(x) &\leq \mathbb{P}[S(m^{-1}(x_\epsilon)) > x-1] \int_0^{m^{-1}(x_\epsilon)} \sum_{i=1}^{\infty} q_i^2 e^{-q_i t}\, dt \\ &= \mathbb{P}[S(m^{-1}(x_\epsilon)) > x-1] \sum_{i=1}^{\infty} q_i (1 - e^{-q_i m^{-1}(x_\epsilon)}) \\ &\leq \mathbb{P}[S(m^{-1}(x_\epsilon)) > x-1]\end{aligned}$$

which, by $m(m^{-1}(x_\epsilon)) = (1-\epsilon)(x-1)$ and Lemma 4, implies

$$I_1(x) \leq H e^{-\theta_\epsilon x} = o\left(\frac{1}{x^{\alpha-1}}\right) \quad \text{as} \quad x \to \infty. \tag{12}$$

Next,

$$\begin{aligned}I_2(x) &= \int_{m^{-1}(x_\epsilon)}^{\infty} \sum_{i=1}^{\infty} q_i^2 e^{-q_i t}\, dt \\ &\leq \sum_{i=1}^{\infty} q_i e^{-q_i m^{-1}(x_\epsilon)} \\ &\leq \frac{1}{m^{-1}(x_\epsilon)} \sum_{i=1}^{\lfloor x \rfloor} q_i m^{-1}(x_\epsilon) e^{-q_i m^{-1}(x_\epsilon)} + \sum_{i=\lfloor x \rfloor + 1}^{\infty} q_i.\end{aligned} \tag{13}$$

Since $\sup_{y \geq 0} y e^{-y} = e^{-1}$ implies $q_i m^{-1}(x_\epsilon) e^{-q_i m^{-1}(x_\epsilon)} \leq e^{-1}$ for all i and $\sum_{i=\lfloor x \rfloor + 1}^{\infty} q_i \leq \int_x^{\infty} H/u^\alpha\, du$, the preceding inequality renders

$$I_2(x) \leq \frac{x e^{-1}}{m^{-1}(x_\epsilon)} + \frac{H}{(\alpha-1)x^{\alpha-1}}. \tag{14}$$

Next, from $q_i \leq H/i^\alpha$, the monotonicity of $1 - e^{-xt}$ and Lemma 3 we derive $m(t) \leq H t^{\frac{1}{\alpha}}$, implying $m^{-1}(x_\epsilon) \geq h x^\alpha$, and, therefore

$$I_2(x) \leq \frac{H}{x^{\alpha-1}},$$

which, in conjunction with (12), proves the result. $\diamond$

Lemma 6 *If* $0 \leq q_i \leq H/i^\alpha$, *then*

$$\sum_{i=1}^{\infty} q_i e^{-q_i t} \leq H t^{-1+\frac{1}{\alpha}}.$$

Proof: Similarly as in the proof of Lemma 5, the claim follows easily from $\sup_{y\geq 0} ye^{-y} = e^{-1}$, the assumption $q_i \leq H/i^\alpha$ and

$$\sum_{i=1}^{\infty} q_i e^{-q_i t} \leq \frac{1}{t} \sum_{i=1}^{\lfloor t^{\frac{1}{\alpha}} \rfloor} q_i t e^{-q_i t} + \sum_{i=\lfloor t^{\frac{1}{\alpha}} \rfloor}^{\infty} q_i,$$

where $\lfloor y \rfloor$ is the integer part of y; we omit the details. ◇

4 Main results

In this section we derive our main results in Theorems 2 and 3. These results fully generalize Theorem 3 of [14] that was proved for the independence reference model. Furthermore, our proving method uses probabilistic and sample path arguments, in contrast to the Tauberian approach used in [14].

4.1 Lower bound

In preparation for our main results, we prove the following lower bound that holds for the entire class of stationary and ergodic modulating request processes, as defined in Section 2.

Proposition 2 *Assume that* $q_i \sim c/i^\alpha$ *as* $i \to \infty$, $\alpha > 1$ *and*

$$K(\alpha) \triangleq \left(1 - \frac{1}{\alpha}\right) \left[\Gamma\left(1 - \frac{1}{\alpha}\right)\right]^\alpha, \tag{15}$$

where Γ *is the Gamma function. Then, as* $x \to \infty$

$$\mathbb{P}[C > x] \gtrsim K(\alpha) \mathbb{P}[R > x].$$

Proof: For any $1 > \epsilon > 0$, let $\{B_i^{-\epsilon}(t), i \geq 1\}$ be a sequence of independent Bernoulli random variables with $\mathbb{P}[B_i^{-\epsilon}(t) = 1] = 1 - e^{-q_i(1-\epsilon)t}$, $S_{-\epsilon}(t) \triangleq \sum_{i=1}^{\infty} B_i^{-\epsilon}(t)$ and $m_{-\epsilon}(t) \triangleq \mathbb{E}S_{-\epsilon}(t) = \sum_{i=1}^{\infty}(1 - e^{-(1-\epsilon)q_i t})$. Note that, using the independence reference model interpretation from the beginning of Subsection 3.2, $S_{-\epsilon}(t)$ represents the number of distinct items requested in interval $(-t(1-\epsilon), 0)$. Therefore, we can assume that $S_{-\epsilon}(t)$ is constructed, on a possibly extended probability space, monotonically non-decreasing in t. We also define

$$\nu(t) \triangleq \max_{1\leq k\leq M} |\hat{\pi}_k - \pi_k| \tag{16}$$

that for all $\omega \in \{\nu(t) \leq \epsilon\}$ and $1 \leq k \leq M$, implies

$$\pi_k(1-\epsilon) \leq \hat{\pi}_k \equiv \hat{\pi}_k(t) \leq \pi_k(1+\epsilon),$$

and, therefore

$$q_i(1-\epsilon) \leq \hat{q}_i \equiv \hat{q}_i(t) \leq q_i(1+\epsilon), \tag{17}$$

for all $i \geq 1$. This further implies that for every $\omega \in \{\nu(t) \leq \epsilon\}$

$$\mathbb{P}_{\sigma_t}[B_j(t;J) = 1] = 1 - e^{-\hat{q}_i t} \geq 1 - e^{-(1-\epsilon)q_i t} = \mathbb{P}[B_i^{-\epsilon}(t) = 1].$$

Therefore, for every $\omega \in \{\nu(t) \leq \epsilon\}$, by Strassen's theorem on stochastic dominance (e.g., see Theorem 2.3.1 of [2]), the total number of distinct items $S(t;J) \equiv S_i(t;J) + B_i(t;J)$ requested in $[-t, 0)$ satisfies

$$\mathbb{P}_{\sigma_t}[S(t;J) > x] \geq \mathbb{P}[S_{-\epsilon}(t) > x]. \tag{18}$$

Then, representation expression (10) and equations (17-18) render for any $g_\epsilon > 0$

$$\begin{aligned}\mathbb{P}[C > x] &\geq \mathbb{E}\int_{g_\epsilon x^\alpha}^\infty \sum_{i=1}^\infty q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t}\mathbb{P}_{\sigma_t}[S(t;J) > x]\, dt \\ &\geq \mathbb{E}\int_{g_\epsilon x^\alpha}^\infty \sum_{i=1}^\infty q_i^{(J_0)} q_i^{(J_{-t})} e^{-q_i(1+\epsilon)t}\mathbb{P}[S_{-\epsilon}(t) > x] 1[\nu(t) \leq \epsilon]\, dt.\end{aligned}$$

Now, using the last expression and monotonicity of $S_{-\epsilon}(t)$ we derive

$$\mathbb{P}[C > x] \geq \mathbb{P}[S_{-\epsilon}(g_\epsilon x^\alpha) > x] \int_{g_\epsilon x^\alpha}^\infty \sum_{i=1}^\infty e^{-q_i(1+\epsilon)t}\mathbb{E}\left[q_i^{(J_0)} q_i^{(J_{-t})} 1[\nu(t) \leq \epsilon]\right]\, dt. \tag{19}$$

The ergodicity of J_t and finiteness of its state space implies that uniformly in k, l and all t large enough ($t \geq t_\epsilon$)

$$\mathbb{P}[\nu(t) \leq \epsilon, J_0 = k, J_{-t} = l] \geq (1-\epsilon)\pi_k \pi_l,$$

which yields for all $i \geq 1$ and t large,

$$\mathbb{E}\left[q_i^{(J_0)} q_i^{(J_{-t})} 1[\nu(t) \leq \epsilon]\right] \geq (1-\epsilon)(q_i)^2. \tag{20}$$

Next, if we choose $g_\epsilon = \frac{(1+2\epsilon)^\alpha}{c(1-\epsilon)[\Gamma(1-\frac{1}{\alpha})]^\alpha}$, then, it is easy to check that, by Lemma 3, $\mathbb{E}m_{-\epsilon}(g_\epsilon x^\alpha) \sim (1+2\epsilon)x$ as $x \to \infty$, from which it follows that for all x large ($x \geq x_\epsilon$) $\mathbb{E}m_{-\epsilon}(g_\epsilon x^\alpha) \geq (1+\epsilon)x$. Therefore, by Lemma 4, for all $x \geq x_\epsilon$

$$\mathbb{P}[S_{-\epsilon}(g_\epsilon x^\alpha) > x] \geq 1 - \epsilon.$$

Thus, replacing the last inequality and (20) in (19), we conclude that for all large x

$$\mathbb{P}[C > x] \geq \frac{(1-\epsilon)^2}{(1+\epsilon)^2}\int_{g_\epsilon x^\alpha}^\infty \sum_{i=1}^\infty (q_i(1+\epsilon))^2 e^{-q_i(1+\epsilon)t}\, dt. \tag{21}$$

In order to estimate the last integral, we observe that, by Lemma 2, for all $t \geq t_\epsilon$

$$\sum_{i=1}^\infty (q_i(1+\epsilon))^2 e^{-q_i(1+\epsilon)t} \geq (1-\epsilon)\frac{((1+\epsilon)c)^{\frac{1}{\alpha}}}{\alpha}\Gamma\left(2 - \frac{1}{\alpha}\right) t^{-2+\frac{1}{\alpha}}.$$

Hence, using the last estimate in (21) and computing the integral result in

$$\mathbb{P}[C > x] \geq \frac{(1-\epsilon)^3}{(1+\epsilon)^2}\frac{((1+\epsilon)c)^{\frac{1}{\alpha}}}{\alpha - 1}\Gamma\left(2 - \frac{1}{\alpha}\right)(g_\epsilon x^\alpha)^{-1+\frac{1}{\alpha}},$$

that, in conjunction with the definition of g_ϵ, for all $x \geq x_\epsilon$ yields

$$\mathbb{P}[C > x] \geq \frac{(1-\epsilon)^{4-\frac{1}{\alpha}}}{(1+2\epsilon)^{1+\alpha-\frac{1}{\alpha}}} K(\alpha) \frac{c}{(\alpha-1)x^{\alpha-1}}.$$

The last bound and the asymptotic behavior of the request distribution $\mathbb{P}[R > x] \sim c/((\alpha-1)x^{\alpha-1})$ further imply

$$\liminf_{x\to\infty} \frac{\mathbb{P}[C > x]}{\mathbb{P}[R > x]} \geq \frac{(1-\epsilon)^{4-\frac{1}{\alpha}}}{(1+2\epsilon)^{1+\alpha-\frac{1}{\alpha}}} K(\alpha),$$

which, by passing $\epsilon \downarrow 0$, concludes the proof. ◇

4.2 General modulation

In this section we prove our first main result, with the underlying process J_t being some general process with sufficiently fast rate of convergence of empirical distribution.

Theorem 2 *If $q_i \sim c/i^\alpha$ as $i \to \infty$, $\alpha > 1$ and for any $\epsilon > 0$*

$$\max_{1\leq k\leq M} \mathbb{P}[|\hat{\pi}_k(t) - \pi_k| > \epsilon] = o\left(t^{\frac{1}{\alpha}-2}\right) \quad as \quad t \to \infty, \tag{22}$$

then

$$\mathbb{P}[C > x] \sim K(\alpha)\mathbb{P}[R > x] \quad as \quad x \to \infty,$$

with $K(\alpha)$ as defined in (15).

Remark 3 *This result and Theorem 3 of the following subsection show that LRU fault probability is asymptotically invariant to the modulating process and behaves the same as in the case of i.i.d. requests with frequencies equal to the marginal distribution q_i. $K(\alpha)$ is monotonically increasing in α with $\lim_{\alpha\to 1} K(\alpha) = 1$ and $\lim_{\alpha\to\infty} K(\alpha) = e^\gamma \approx 1.78$, where γ is the Euler constant; this was formally provided in Theorem 3 of [14].*

Remark 4 *Condition (22) may exclude the set of processes whose autocorrelation functions decay slower than $t^{\frac{1}{\alpha}-2}$, in particular long-range dependent modulating processes J_t. To see this, consider the case when the jump points $(T_n - T_{n-1})$ are i.i.d. with the first jump point distributed as*

$$\mathbb{P}[T_1 > t] = \frac{d}{t^\beta}, \quad 0 < \beta \leq 1$$

for some constant d, and the modulating process observed at this points J_{T_n} is a finite state Markov chain independent of $\{T_n\}$. Then, Theorem 7 of [16], shows that the autocorrelation function of J_t satisfies

$$\rho(t) \sim \mathbb{P}[T_1 > t] \quad as \quad t \to \infty,$$

implying that $\int_1^\infty \rho(t)\, dt = \infty$; hence, J_t is long-range dependent. On the other hand, since J_0 is independent of T_1

$$\begin{aligned}\mathbb{P}[|\hat{\pi}_k(t) - \pi_k| > \epsilon] &\geq \mathbb{P}[|\hat{\pi}_k(t) - \pi_k| > \epsilon, T_1 > t] \\ &= \mathbb{P}[|1[J_0 = k] - \pi_k| > \epsilon]\mathbb{P}[T_1 > t] = \frac{d_1}{t^\beta},\end{aligned}$$

where $d_1 = d\mathbb{P}[|1[J_0 = k] - \pi_k| > \epsilon]$. *Therefore,*

$$t^{2-\frac{1}{\alpha}}\mathbb{P}[|\hat{\pi}_k - \pi_k| > \epsilon] \geq d_1 t^{2-\frac{1}{\alpha}-\beta} \to \infty \;\; as \;\; t \to \infty,$$

which violates condition (22).

Proof: Using $S(t;J) \equiv S_i(t;J) + B_i(t;J) \geq S_i(t;J)$ and the representation in (10), for any $h > 0$

$$\begin{aligned}\mathbb{P}[C > x] &\leq \mathbb{E}\int_0^{hx^\alpha} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J) > x-1]\,dt + \mathbb{E}\int_{hx^\alpha}^{\infty} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J) > x-1]\,dt \\ &\triangleq I_1(x) + I_2(x), \end{aligned} \tag{23}$$

where

$$\hat{f}(t) \triangleq \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t} \leq \sum_{i=1}^{\infty} q_i^{(J_0)} = 1. \tag{24}$$

Furthermore, the empirical distributions are uniformly bounded by $\hat{q}_i = \sum_{k=1}^M \hat{\pi}_k q_i^{(k)} \leq \sum_{k=1}^M q_i^{(k)} \leq \bar{q}_i \triangleq q_i / \min_k \pi_k < \infty$, since $\min_k \pi_k > 0$. Then, we define a sequence of independent Bernoulli random variables $\{\bar{B}_i(t), i \geq 1\}$, with $\mathbb{P}[\bar{B}_i(t) = 1] = 1 - e^{-\bar{q}_i t}$ and $\bar{S}(t) = \sum_{i=1}^{\infty} \bar{B}_i(t)$; similarly as in the proof of the lower bound, $\bar{S}(t)$ can be constructed non-decreasing in t. Note that for every ω, $\mathbb{P}_{\sigma_t}[B_i(t;J) = 1] \leq \mathbb{P}[\bar{B}_i(t) = 1]$ and, therefore, by stochastic dominance (see Strassen's Theorem 2.3.1 in [2]), we obtain $\mathbb{P}_{\sigma_t}[S(t;J) > x-1] \leq \mathbb{P}[\bar{S}(t) > x-1]$ uniformly in ω. Using this observation and the monotonicity of $\bar{S}(t)$, we arrive at

$$I_1(x) \leq \int_0^{hx^\alpha} \mathbb{P}[\bar{S}(t) > x-1]\,dt \leq Hx^\alpha \mathbb{P}[\bar{S}(hx^\alpha) > x-1]. \tag{25}$$

Now, due to Lemma 3, $\mathbb{E}\bar{S}(t) \leq Ht^{\frac{1}{\alpha}}$, and therefore, we can always find h small enough such that for any $\epsilon > 0$ and all x large enough

$$\mathbb{E}\bar{S}(hx^\alpha) < (1-\epsilon)(x-1). \tag{26}$$

Then, (25), (26), Lemma 4 and Lemma 3 imply, as $x \to \infty$

$$I_1(x) \leq Hx^\alpha e^{-\theta_\epsilon x} = o\left(\frac{1}{x^{\alpha-1}}\right). \tag{27}$$

Then, by using $\nu(t)$ as defined in (16), we obtain

$$\begin{aligned} I_2(x) &\leq \mathbb{E}\int_{hx^\alpha}^{\infty} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J) > x-1]\,dt \\ &= \mathbb{E}\int_{hx^\alpha}^{\infty} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J) > x-1]1[\nu(t) \leq \epsilon]\,dt + \mathbb{E}\int_{hx^\alpha}^{\infty} \hat{f}(t)\mathbb{P}_{\sigma_t}[S(t;J) > x-1]1[\nu(t) > \epsilon]\,dt \\ &= I_{21}(x) + I_{22}(x). \end{aligned} \tag{28}$$

Note that, by assumption of the theorem, for any $\delta > 0$ and t large enough, $\mathbb{P}[\nu(t) > \epsilon] \leq \frac{\delta}{t^{2-\frac{1}{\alpha}}}$ and, therefore, using (24), for all x large enough

$$I_{22}(x) \leq \int_{hx^\alpha}^{\infty} \frac{\delta}{t^{2-\frac{1}{\alpha}}}\,dt \leq \frac{\delta}{(1-\frac{1}{\alpha})h^{1-\frac{1}{\alpha}}x^{\alpha-1}}.$$

Thus, since δ can be arbitrarily small

$$I_{22}(x) = o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x \to \infty. \tag{29}$$

Next, we will provide the estimate for $I_{21}(x)$. Similarly as in the proof of the lower bound, we define $S_\epsilon(t) \triangleq \sum_{i=1}^{\infty} B_i^\epsilon(t)$, where $\{B_i^\epsilon(t), i \geq 1\}$ is a sequence of independent Bernoulli random variables with $\mathbb{P}[B_i^\epsilon(t) = 1] = 1 - e^{-q_i(1+\epsilon)t}$. As before, $S_\epsilon(t)$ can be constructed non-decreasing in t. Therefore, by stochastic dominance, for every $\omega \in \{\nu(t) \leq \epsilon\}$,

$$\mathbb{P}_{\sigma_t}[S(t;J) > x-1] \leq \mathbb{P}[S_\epsilon(t) > x-1].$$

Furthermore, since for all ω in $\{\nu(t) \leq \epsilon\}$ (17) holds, by using (24), we obtain that for any constant $g_\epsilon > 0$

$$\begin{aligned} I_{21}(x) &\leq \mathbb{E}\int_0^\infty \sum_{i=1}^{\infty} q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t} \mathbb{P}[S_\epsilon(t) > x-1] 1[\nu(t) \leq \epsilon]\, dt \\ &\leq \mathbb{E}\int_0^{g_\epsilon x^\alpha} \mathbb{P}[S_\epsilon(t) > x-1]\, dt + \int_{g_\epsilon x^\alpha}^\infty \sum_{i=1}^{\infty} \mathbb{E}\left[q_i^{(J_0)} q_i^{(J_{-t})} 1[\nu(t) \leq \epsilon]\right] e^{-(1-\epsilon)q_i t}\, dt. \end{aligned} \tag{30}$$

If we select $g_\epsilon = \frac{(1-2\epsilon)^\alpha}{c(1+\epsilon)[\Gamma(1-\frac{1}{\alpha})]^\alpha}$, then, due to Lemma 3, $\mathbb{E}S_\epsilon(g_\epsilon x^\alpha) \sim (1-2\epsilon)x$, which implies that for all x large enough ($x \geq x_\epsilon$),

$$\mathbb{E}S_\epsilon(g_\epsilon x^\alpha) < (1-\epsilon)(x-1).$$

Hence, since $S_\epsilon(t)$ is non-decreasing, by applying Lemmas 4 and 3 we conclude that for x large ($x \geq x_\epsilon$)

$$\int_0^{g_\epsilon x^\alpha} \mathbb{P}[S_\epsilon(t) > x-1]\, dt \leq g_\epsilon x^\alpha \mathbb{P}[S_\epsilon(g_\epsilon x^\alpha) > x-1] \leq H x^\alpha e^{-\theta_\epsilon x} = o\left(\frac{1}{x^{\alpha-1}}\right). \tag{31}$$

At this point, it remains to derive an estimate of the second integral in (30). Similarly as in the proof of the lower bound, since J_t is ergodic and has finitely many states, for all $i \geq 1$ and t large ($t \geq t_\epsilon$)

$$\mathbb{E}[q_i^{(J_0)} q_i^{(J_{-t})} 1[\nu(t) \leq \epsilon]] \leq (1+\epsilon)(q_i)^2.$$

This implies that for x large enough ($x \geq x_\epsilon$), the second term in (30) is bounded by

$$\frac{1+\epsilon}{(1-\epsilon)^2} \int_{g_\epsilon x^\alpha}^\infty \sum_{i=1}^{\infty} ((1-\epsilon)q_i)^2 e^{-(1-\epsilon)q_i t}\, dt.$$

Bounding the preceding expression is analogous to evaluating the integral in (21), i.e., we use Lemma 2 to upper bound the sum under the integral for large x and then compute the integral for the chosen g_ϵ. Therefore, combining the bound obtained in this way with (31), (30), (29), (28),(27) and (23), we derive

$$\limsup_{x\to\infty} \mathbb{P}[C > x] x^{\alpha-1} \leq \frac{(1+\epsilon)^{3-\frac{1}{\alpha}}}{(1-2\epsilon)^{1+\alpha-\frac{1}{\alpha}}} K(\alpha) \frac{c}{(\alpha-1)},$$

which, by passing $\epsilon \downarrow 0$, finishes the proof.

$\diamond$

4.3 Semi-Markov modulation

In order to cover the cases when condition (22) is not satisfied, e.g. those examples from Remark 4 that exhibit the long-range dependence, we assume the following detailed structure of the modulating process. We consider the class of semi-Markov processes, that is uniquely defined by the following evolution of J_t at jump points T_n

$$\begin{aligned}\mathbb{P}[J_{T_n} = k, T_{n+1} - T_n \le t | J_{T_j}, T_{j+1}, j < n] \\ = p_{ik}(1 - F_k(t)) \quad \text{on} \quad \{J_{T_{n-1}} = i\},\end{aligned}$$

where $F_k(t) = \mathbb{P}[T_{n+1} - T_n \le t | J_{T_n} = k]$ and $p_{ik} = \mathbb{P}[J_{T_{n+1}} = k | J_{T_n} = i]$ (see Section 5 of Chapter 10 in [7]). We assume that $\{p_{ij}\}$ is a stationary and ergodic (irreducible) finite state Markov chain. Recall that, without loss of generality, we set $T_0 < 0 \le T_1$. Throughout this section we assume that $\mathbb{E}[T_2 - T_1]^{1+\delta} < \infty$ and define

$$\nu_k \triangleq \mathbb{P}[J_{T_1} = k], \qquad \mu_k \triangleq \mathbb{E}[T_2 - T_1 | J_{T_1} = k] \qquad \text{and} \qquad \mu \triangleq \mathbb{E}[T_2 - T_1] = \sum_{k=1}^{M} \nu_k \mu_k.$$

Then, for J_t in stationarity, $\mathbb{E}(T_2 - T_1)^{1+\delta} < \infty$ implies for all $1 \le k \le M$

$$\mathbb{P}[T_1 > x | J_0 = k] = \frac{\int_x^\infty (1 - F_k(u))\, du}{\mu_k} = o\left(\frac{1}{x^\delta}\right) \qquad \text{as} \qquad x \to \infty. \tag{32}$$

For J_t as described above, we state our second main result.

Theorem 3 *Assume that J_t is semi-Markov with $\mathbb{E}[T_2 - T_1]^{1+\delta} < \infty$, for some $\delta > 0$. If $q_i \sim c/i^\alpha$ as $i \to \infty$, $\alpha > 1$, then*

$$\mathbb{P}[C > x] \sim K(\alpha)\mathbb{P}[R > x] \qquad \textit{as} \qquad x \to \infty,$$

with $K(\alpha)$ as defined in (15).

In preparation for the proof we define the moments of reversed jump points $T_n^r \triangleq -T_{-n}$, $n \ge 0$; this notation is convenient since C depends on J_t for negative values of $t \le 0$. Note that (T_0^r, J_0) is equal in distribution to (T_1, J_0) and, thus, (32) holds for T_0^r as well.

Heuristic outline of the proof: The proof of the lower bound follows from Proposition 2. Hence, in order to complete the proof, we need to prove the upper bound. To this end, we observe that $\hat{f}(t)$, as defined in (24), is a random variable measurable with respect to σ_t. Therefore, using $S(t; J) \ge S_i(t; J)$ and $\mathbb{P}_{\sigma_t}[S(t; J) > x] = \mathbb{E}_{\sigma_t} 1[S(t; J) > x]$, the integral representation in (10) is bounded by

$$\begin{aligned}\mathbb{P}[C > x] &\le \mathbb{E} \int_0^\infty \hat{f}(t) 1[S(t; J) > x - 1]\, dt \\ &= \mathbb{E} \int_0^{T_0^r} + \mathbb{E} \int_{T_0^r}^{T^r_{\lfloor x^{1/3} \rfloor}} + \mathbb{E} \int_{T^r_{\lfloor x^{1/3} \rfloor}}^\infty \\ &\triangleq I_1(x) + I_2(x) + I_3(x).\end{aligned} \tag{33}$$

For a given initial state $J_0 = k$, the integral representation in $I_1(x)$ approximately corresponds to the case of i.i.d. requests, represented in (11), where q_i is replaced by $q_i^{(k)}$ and the integration is truncated by a random

time T_0^r. Now, using the fact that $\mathbb{P}[T_0^r > x] \to 0$ as $x \to \infty$ and Lemma 5, we estimate $I_1(x) = o(1/x^{\alpha-1})$ as $x \to \infty$. Next, observe that for x large enough $T^r_{\lfloor x^{1/3} \rfloor} \approx x^{1/3}\mu$. Then, by using $\hat{f}(t) \leq 1$ and the definition of $\bar{S}(t)$ from the proof of Theorem 2, we conclude

$$\begin{aligned} I_2(x) &\lessapprox \int_0^{x^{1/3}\mu} \mathbb{P}[\bar{S}(t) > x-1]\,dt \\ &\leq x^{1/3}\mu \mathbb{P}[\bar{S}(x^{1/3}\mu) > x-1] \\ &= o\left(\frac{1}{x^{\alpha-1}}\right) \text{ as } x \to \infty, \end{aligned}$$

where in the last equality we exploited Lemma 4. Finally, due to ergodicity of the process J_t, for t large enough $\hat{q}_i \approx q_i$ and, therefore, from the definitions of $B_i(t;J)$ and $S(t;J)$, we deduce that $S(t;J) \approx S(t)$, where $S(t)$ coresponds to the number of distinct requests in $[-t,0)$ for the case of an i.i.d. requests with distribution q_i, as defined in Subsection 3.2. Hence, for x large enough, $I_3(x)$ is approximately

$$\begin{aligned} I_3(x) &\approx \mathbb{E}\int_{x^{1/3}\mu}^{\infty} \hat{f}(t) 1[S(t;J) > x-1]\,dt \\ &\approx \int_{x^{1/3}\mu}^{\infty} \sum_{i=1}^{\infty} e^{-q_i t}\mathbb{E}[q_i^{(J_0)} q_i^{(J_{-t})}]\mathbb{P}[S(t) > x-1]\,dt \\ &\lessapprox \int_0^{\infty} \sum_{i=1}^{\infty} e^{-q_i t}(q_i)^2 \mathbb{P}[S_i(t) > x-2]\,dt, \end{aligned}$$

since, by (1), $\mathbb{E}[q_i^{(J_0)} q_i^{(J_{-t})}] \approx (q_i)^2$ and $S_i(t) \geq S(t) - 1$. The last expression equals to the case of i.i.d. requests stated in equation (11) and can be estimated using either Theorem 3 of [14] or our Theorem 2. A rigorous proof of the theorem is much more involved and very technical and, therefore, we present it in separate Section 7 of this paper. $\diamond$

5 Numerical examples

In this subsection, we provide three simulation experiments that illustrate Theorems 2 and 3. We consider the case where the underlying process J_t is a two-state $\{0,1\}$ semi-Markov process with parameters implying strong correlation. Since the asymptotic results were obtained first by passing the list size N to infinity and then investigating the tail of the limiting search cost distribution, it can be expected that the asymptotic expression gives a reasonable approximation for $\mathbb{P}[C^N > k]$ when both N and k are large. However, it is surprising how accurately the approximation works even for relatively small values of N and almost all values of $k < N$.

The initial position of items in the list is chosen uniformly at random. In each experiment, before we conduct measurements, we allow 10^7 units of warm-up time for the system to reach its steady state. In order to make sure that the simulation results are accurate, we run another simulation, where we take the initial position of items according to the reversed order of their popularity (i.e. the first item is the least popular, etc.). In all these experiments, the measured results are almost identical for these different initial conditions. Therefore, one can safely assume that the experiments reached the steady state. The actual measurement time is set to be 10^7 units long. In all of the experiments, the measurements are conducted for

cache sizes $k = j * 50, 1 \leq j \leq 16$ and are presented with star "*" symbols on Figures 1, 2 and 3, while our approximation, $K(\alpha)\mathbb{P}[R > n]$, is represented with the solid line on the same figures.

The total number of documents in all three experiments is set to $N = 1000$. The Markovian transitions of the two-state modulating process are $p_{01} = p_{10} = 1$. We use τ^0 and τ^1 to denote the variables equal in distribution to the sojourn times corresponding to states 0 and 1, respectively. In the first two experiments τ^0 and τ^1 are discrete random variables, while in the third experiment they are continuous.

Example 1 *In this experiment we choose discrete random variables τ^0 and τ^1 to be distributed as $\mathbb{P}[\tau^1 = i * 10] = \mathbb{P}[\tau^0 = i * 10] = a * (1/(i * 10)^3 - 1/((i + 1) * 10)^3)$, where $i \in \{1, \ldots, 10^4\}$ and $a = 10^3(1 - 1/(10^4 + 1)^3)^{-1}$. In state 0, only odd items are requested according to $q^{(0)}_{2i+1} = H^0_N/(2i + 1)^{1.4}$, $q^{(0)}_{2i} = 0$, where $1/H^0_N = \sum_{i=0}^{499} 1/(2i+1)^{1.4}$, while in state 1, the probabilities are concentrated exclusively on even documents, $q^{(1)}_{2i} = H^1_N/(2i)^{1.4}$, $q^{(1)}_{2i+1} = 0$, where $1/H^1_N = \sum_{i=1}^{N/2} 1/(2i)^{1.4}$. The experimental results are presented in Figure 1. This model corresponds to the case where two different classes of clients request documents from disjoint sets. Even in this extreme scenario, our approximation $K(\alpha)\mathbb{P}[R > k]$ matches very precisely the simulated results.*

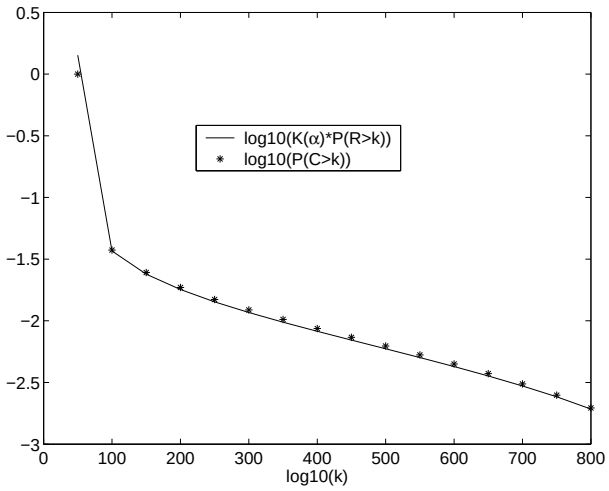

Figure 1: Illustration for Example 1

Example 2 *Here, we select variables τ^0 and τ^1 to be distributed as $\mathbb{P}[\tau^1 = i * 10] = \mathbb{P}[\tau^0 = i * 10] = b * (1/(i * 10)^{0.8} - 1/((i + 1) * 10)^{0.8})$, where $i \in \{1, \ldots, 10^4\}$ and $b = 10^{0.8}(1 - 1/(10^4 + 1)^{0.8})^{-1}$. In state 0, items are requested according to distribution $q^{(0)}_i = C^0_N/i^{1.4}$, where $1/C^0_N = \sum_{i=1}^{N} 1/i^{1.4}$, and in state 1, the popularity of documents is given by $q^{(1)}_i = C^1_N/i^{4.0}$, where $1/C^1_N = \sum_{i=1}^{N} 1/i^4$. Our intention in this experiment is to show that only the heavier tailed request process impacts the LRU performance. The simulation results in this case are presented in Figure 2. As in the preceding experiment, we obtain accurate agreement between the approximation and simulation.*

Example 3 *Now, we illustrate the case where $\mathbb{P}[\tau^1 > t] = e^{-3t}$, $t \in [0, \infty)$ (exponential distribution) and $\mathbb{P}[\tau^0 > t] = 1/t^{0.8}$, $t \in [1, 10^5]$ and $\mathbb{P}[\tau^0 > t] = 0$ for $t > 10^5$. In state 0, items are requested according to distribution $q^{(0)}_i = H^0_N/i^3$, where $1/H^0_N = \sum_{i=1}^{N} 1/i^3$. In state 1, the popularity of documents is*

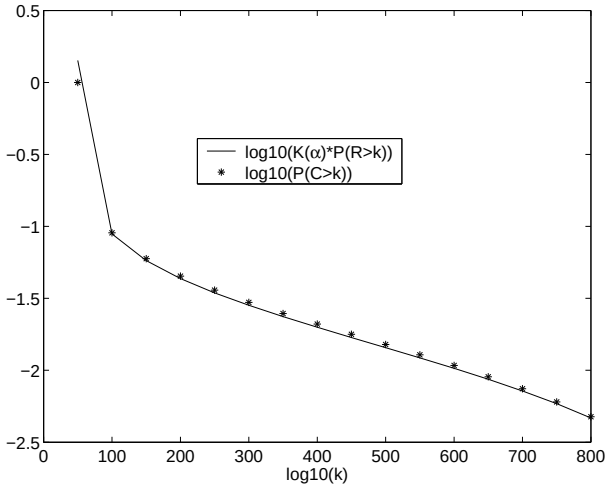

Figure 2: Illustration for Example 2

$q_i^{(1)} = H_N^1/i^{1.4}$, *where* $1/H_N^1 = \sum_{i=1}^N 1/i^{1.4}$. *This experiment shows that even in the case when* $\mathbb{E}\tau^0 = 46 \gg \mathbb{E}\tau^1 = 1/3$, *the tail of the search cost distribution is asymptotically dominated by the heavier tail of requests in state* 1. *Again, the excellent agreement of the approximation with simulated results is apparent from Figure 3.*

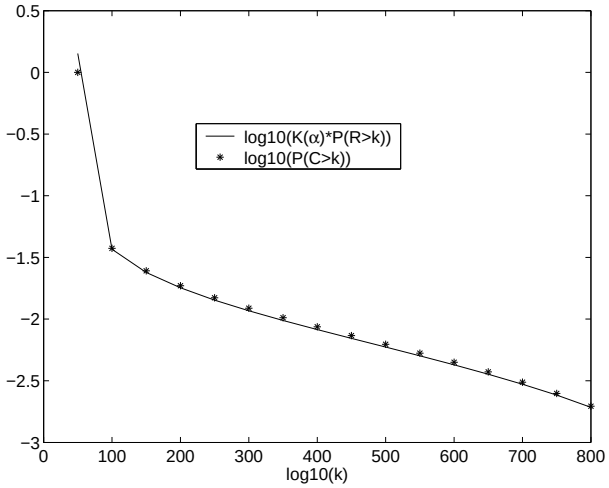

Figure 3: Illustration for Example 3

6 Concluding remarks

In this paper we investigated the asymptotic behavior of the LRU cache fault probability, or equivalently the MTF search cost distribution, for a class of semi-Markov modulated request processes. This class of processes provides both the analytical tractability and flexibility of modeling a wide range of statistical correlation, including the empirically measured long-range dependence. When the marginal probability mass

function of requests follows Zipf's law, our main results show that the LRU fault probability is asymptotically proportional to the tail of the request distribution. These results assume the same form as the recently developed asymptotics for the i.i.d. requests, implying that the LRU cache fault probability is invariant to the underlying, possibly strong, dependency structure in the document request sequence. This surprising insensitivity suggests that one may not need to model accurately, if at all, the statistical correlation in the request sequence. Hence, this may simplify the modeling process of the Web access patterns and further improve the speed of simulating network caching systems.

Our results are further validated using simulation. The excellent agreement between the analytical and experimental results implies the potential use of our approximation in predicting the performance and properly engineering Web caches. The explicit nature, high degree of accuracy and low computational complexity of our result contrast the lengthy procedure of simulation experiments.

7 Proof of Theorem 3

In order to prove the theorem we will need the following technical lemma. Recall the definition of T_n^r from Subsection 4.3.

Lemma 7 *If $\mathbb{E}[T_1^r - T_0^r]^{1+\delta} < \infty$, then $h > 0$ such that uniformly for all $n \leq hx$*

$$\mathbb{P}[T_n^r - T_0^r > x] \leq no\left(\frac{1}{x^{1+\delta}}\right) \quad as \quad x \to \infty.$$

Proof of Lemma 7: We construct a sequence $\{X_i, i \leq n\}$ of i.i.d. random variables with

$$\bar{F}(x) \triangleq \mathbb{P}[X_i > x] = \max_{1 \leq k \leq M} \bar{F}_k(x),$$

where $\bar{F}_k(x)$, $1 \leq k \leq M$ is defined at the beginning of Section 4. Therefore, $\mathbb{P}[X_i > x] \geq \mathbb{P}[T_i^r - T_{i-1}^r > x | J_{T_i^r}]$ and

$$\begin{aligned}
\mathbb{P}[T_n^r - T_0^r > x] &= \mathbb{P}\left[\sum_{i=1}^{n}(T_i^r - T_{i-1}^r) > x\right] \\
&= \mathbb{E}\left[\mathbb{P}\left[\sum_{i=1}^{n}(T_i^r - T_{i-1}^r) > x \,\middle|\, J_{T_j}, 1 \leq j \leq n\right]\right] \\
&\leq \mathbb{P}\left[\sum_{i=1}^{n} X_i > x\right] = \mathbb{P}\left[\sum_{i=1}^{n} X_i - n\mathbb{E}X_i > x - n\mathbb{E}X_i\right].
\end{aligned}$$

Now, since $\mathbb{E}[T_1^r - T_0^r]^{1+\delta} < \infty$, we conclude $\mathbb{E}X_i^{1+s} < H$, for any $0 \leq s \leq \delta$ and some sufficiently large positive constant H and therefore, uniformly for all $n \leq hx$, we obtain

$$\mathbb{P}[T_n^r - T_0^r > x] \leq \mathbb{P}\left[\sum_{i=1}^{n} X_i - n\mathbb{E}X_i > x - hHx\right].$$

Now, by taking $h > 0$ such that $hH = 1/2$ and applying Corollary 1.6 of [19] we conclude the proof. $\diamond$

Proof of the theorem: In view of the heuristic outline of the proof from Subsection 4.3, we proceed by

deriving the upper bounds for the expressions $I_j(x)$ defined in (33). In order to estimate $I_1(x)$, we first condition on T_0^r being respectively greater and smaller than hx^α

$$
\begin{aligned}
I_1(x) &= \mathbb{E}\int_0^{T_0^r}\sum_{i=1}^{\infty}(q_i^{(J_0)})^2 e^{-q_i^{(J_0)}t}1[S(t;J)>x-1]\,dt \\
&\le \mathbb{E}\int_0^{T_0^r}\sum_{i=1}^{\infty}(q_i^{(J_0)})^2 e^{-q_i^{(J_0)}t}1[S(t;J)>x-1]1[T_0^r>hx^\alpha]\,dt + \mathbb{E}\int_0^{T_0^r}1[S(t;J)>x-1]1[T_0^r\le hx^\alpha]\,dt \\
&\triangleq I_{11}(x)+I_{12}(x).
\end{aligned}
$$

Next, we define $S_i^{(k)}(t)\triangleq\sum_{j\neq i}B_j^{(k)}(t)$, $S^{(k)}(t)=S_i^{(k)}(t)+B_i^{(k)}(t)$, where $\{B_i^{(k)}(t), i\ge 1\}$ is a sequence of independent Bernoulli random variables with $\mathbb{P}[B_i^{(k)}(t)=1]=1-e^{-q_i^{(k)}t}$. Then, from the definition of $S(t;J)$ it follows that $\mathbb{P}[S(t;J)>x|J_0=k,t<T_0^r]=\mathbb{P}[S^{(k)}(t)>x]$. Thus, using this fact, $q_i^{(k)}\le\bar{q}_i\le H/i^\alpha$, Lemma 5, and equation (32), we obtain

$$
\begin{aligned}
I_{11}(x) &\le \sum_{k=1}^{M}\mathbb{P}[J_0=k,T_0^r>hx^\alpha]\int_0^\infty\sum_{i=1}^{\infty}(q_i^{(k)})^2e^{-q_i^{(k)}t}\mathbb{P}[S_i^{(k)}(t)>x-2]\,dt \\
&\le M\mathbb{P}[T_0^r>hx^\alpha]\frac{H}{x^{\alpha-1}} = o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x\to\infty. \qquad (34)
\end{aligned}
$$

In estimating $I_{12}(x)$ we use $T_0^r\le hx^\alpha$ and exactly the same arguments as in (25 - 27), rendering

$$
I_{12}(x)\le hx^\alpha\mathbb{P}[\bar{S}(hx^\alpha)>x-1]=o\left(\frac{1}{x^{\alpha-1}}\right)\ \text{ as }\ x\to\infty.
$$

Thus, the preceding bound and (34) imply

$$
I_1(x)=o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x\to\infty. \qquad (35)
$$

At this point, we provide an estimate for $I_2(x)$. If we define $I_n^*(x)\triangleq\mathbb{E}\int_{T_{n-1}^r}^{T_n^r}\hat{f}(t)1[S(T_n^r;J)>x-1]\,dt$, then

$$
\begin{aligned}
I_n^*(x) &= \mathbb{E}\int_{T_{n-1}^r}^{T_n^r}\hat{f}(t)1[T_0^r>hx^\alpha]1[S(T_n^r;J)>x-1]\,dt + \mathbb{E}\int_{T_{n-1}^r}^{T_n^r}\hat{f}(t)1[T_0^r\le hx^\alpha]1[S(T_n^r;J)>x-1]\,dt \\
&\triangleq I_{n,1}^*(x)+I_{n,2}^*(x).
\end{aligned}
$$

Next, when we replace the bound

$$
\hat{f}(t)\le\sum_{i=1}^{\infty}q_i^{(J_0)}q_i^{(J_{T_n^r})}e^{-q_i^{(J_0)}T_0^r}e^{-q_i^{(J_{T_n^r})}(t-T_{n-1}^r)}\ \text{ for }\ t\in(T_{n-1}^r,T_n^r] \qquad (36)
$$

in $I_{n,1}^*(x)$ and evaluate the integral, we derive

$$
I_{n,1}^*(x)\le\mathbb{E}\sum_{i=1}^{\infty}1[T_0^r>hx^\alpha]q_i^{(J_0)}e^{-q_i^{(J_0)}hx^\alpha}(1-e^{-q_i^{(J_{T_n^r})}(T_n^r-T_{n-1}^r)}).
$$

Then, using $1-e^{-x} \le x$ for $x \ge 0$, $\max(q_i^{(J_0)}, q_i^{(J_{T_n^r})}) \le \bar{q}_i \le H/i^\alpha$ and $\sup_{y\ge 0} ye^{-y} = e^{-1}$, similarly as in (13), we arrive at

$$\begin{aligned} I^*_{n,1}(x) &\le \mathbb{E}\sum_{i=1}^{\lfloor x \rfloor} \frac{e^{-1}1[T_0^r > hx^\alpha]}{hx^\alpha}\bar{q}_i(T_n^r - T_{n-1}^r) + \mathbb{E}\sum_{i=\lfloor x \rfloor}^{\infty} 1[T_0^r > hx^\alpha](\bar{q}_i)^2(T_n^r - T_{n-1}^r) \\ &\le \mathbb{E}[(T_n^r - T_{n-1}^r)1[T_0^r > hx^\alpha]]\frac{H}{x^\alpha}. \end{aligned} \tag{37}$$

Since random variables $(T_n^r - T_{n-1}^r)$, $n \ge 1$ and $1[T_0^r > hx^\alpha]$ are conditionally independent given J_0, we derive

$$\begin{aligned} \mathbb{E}[(T_n^r - T_{n-1}^r)1[T_0^r > hx^\alpha]] &= \mathbb{E}[\mathbb{E}[T_n^r - T_{n-1}^r|J_0]\mathbb{P}[T_0^r > hx^\alpha|J_0]] \\ &\le (\max_{1\le k\le M} \mathbb{E}[T_n^r - T_{n-1}^r|J_0 = k])\mathbb{P}[T_0^r > hx^\alpha] \\ &\le (\max_{1\le k\le M} \mu_k)\mathbb{P}[T_0^r > hx^\alpha]. \end{aligned} \tag{38}$$

Thus, the proceeding bound, (37) and (32) yield uniformly in n

$$I^*_{n,1}(x) = o\left(\frac{1}{x^\alpha}\right) \qquad \text{as} \qquad x \to \infty. \tag{39}$$

Next, we compute an estimate for $I^*_{n,2}(x)$. Using (36) and computing the integral in $I^*_{n,2}(x)$ result in

$$\begin{aligned} I^*_{n,2}(x) &\le \mathbb{E}\sum_{i=1}^{\infty} 1[T_0^r \le hx^\alpha, S(T_n^r; J) > x-1]q_i^{(J_0)}e^{-q_i^{(J_0)}T_0^r}(1 - e^{-q_i^{(J_{T_n^r})}(T_n^r - T_{n-1}^r)}) \\ &\le \mathbb{P}[T_0^r \le hx^\alpha, S(T_n^r; J) > x-1] \\ &\le \mathbb{P}[S(T_n^r; J) > x-1, T_0^r \le hx^\alpha, S(T_0^r; J) \le \epsilon x - 1] + \mathbb{P}[T_0^r \le hx^\alpha, \bar{S}(T_0^r) > \epsilon x - 1], \end{aligned} \tag{40}$$

since $\bar{S}(t)$ stochastically dominates $S(t; J)$. Let $S(u, t; J)$, $0 < u < t$, be the number of distinct items requested in $[-t, -u)$; then, it is easy to see that $S(T_n^r; J) \le S(T_0^r; J) + S(T_0^r, T_n^r; J)$. Thus, if we choose h small enough such that $\mathbb{E}\bar{S}(hx^\alpha) \le \frac{\epsilon}{2}x$ for large x, then, by Lemma 4, we obtain ($x \ge x_\epsilon$)

$$\begin{aligned} I^*_{n,2}(x) &\le \mathbb{P}[S(T_0^r, T_n^r; J) > (1-\epsilon)x] + \mathbb{P}[\bar{S}(hx^\alpha) > \epsilon x - 1] \\ &\le \mathbb{P}[\bar{S}(T_n^r - T_0^r) > (1-\epsilon)x] + He^{-\theta_\epsilon x}. \end{aligned} \tag{41}$$

Now, if we pick h small enough such that $\mathbb{E}\bar{S}(hx^\alpha) < (1-2\epsilon)x$, we obtain that uniformly for all $n \le x^{\frac{1}{3}}$

$$\begin{aligned} \mathbb{P}[\bar{S}(T_n^r - T_0^r) > (1-\epsilon)x] &\le \mathbb{P}[\bar{S}(hx^\alpha) > (1-\epsilon)x] + \mathbb{P}[T_n^r - T_0^r > hx^\alpha] \\ &\le \mathbb{P}[\bar{S}(hx^\alpha) > (1-\epsilon)x] + n\mathbb{P}\left[T_1^r - T_0^r > \frac{hx^\alpha}{n}\right] \\ &\le He^{-\theta_\epsilon x} + o\left(\frac{1}{x^{\alpha-2/3}}\right) \qquad \text{as} \qquad x \to \infty, \end{aligned} \tag{42}$$

where in the second inequality we used the union bound, and in the last expression Lemma 4 and $\mathbb{E}(T_1^r - T_0^r) < \infty$. This implies $I^*_{n,2}(x) = o\left(\frac{1}{x^{\alpha-2/3}}\right)$ as $x \to \infty$, uniformly for all $n \le x^{1/3}$. Therefore, in conjuction with (39), we derive

$$I_2(x) = \sum_{n=1}^{\lfloor x^{1/3} \rfloor} I^*_n(x) = o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x \to \infty. \tag{43}$$

In order to derive asymptotics for $I_3(x)$, we use the fact that semi-Markov process J_t observed at jump points $\{J_{T_n^r}, -\infty < n < \infty\}$ is a Markov chain. Define the amount of time that J_t spends in state k in $T_{0,n}^r \equiv (T_0^r, T_n^r)$ as

$$\tau_k(T_{0,n}^r) \triangleq \sum_{i=0}^{n-1} 1[J_{T_{i+1}^r} = k](T_{i+1}^r - T_i^r),$$

Then,

$$\mathbb{E}\tau_k(T_{0,n}^r) = \sum_{i=0}^{n-1} \mathbb{P}[J_{T_{i+1}^r} = k]\mathbb{E}[T_{i+1}^r - T_i^r | J_{T_{i+1}^r} = k] = n\nu_k\mu_k = n\mu\pi_k.$$

For any $1 \leq k \leq M$, well-known result on finite state ergodic Markov chains shows that for any $\epsilon > 0$, there is $\theta_{k,\epsilon} > 0$, such that

$$\mathbb{P}\left[\left|\sum_{i=0}^{n-1} 1[J_{T_{i+1}^r} = k] - \nu_k n\right| > \epsilon n\right] \leq e^{-\theta_{k,\epsilon} n}.$$

Next, since intervals $T_i^r - T_{i-1}^r$ are nonnegative random variables with finite mean, they have bounded left tails and, thus, large deviation arguments and the preceeding inequality easily yield for any $\epsilon > 0$ and some $\theta_\epsilon > 0$

$$\mathbb{P}[\tau_k(T_{0,n}^r) < (1-\epsilon)\mathbb{E}\tau_k(T_{0,n}^r)] \leq e^{-\theta_\epsilon n}. \tag{44}$$

Therefore, the probability of the complement of the following set

$$\mathcal{A}(n) \triangleq \cap_{1\leq k\leq M}\{\tau_k(T_{0,n}^r) \geq (1-\epsilon)n\mu_k\nu_k\}$$

is exponentially bounded $\mathbb{P}[\mathcal{A}^c(n)] \leq Me^{-\theta_\epsilon n}$. Then, we decompose $I_3(x)$ as

$$I_3(x) \leq \mathbb{E}\sum_{n=\lfloor x^{1/3}\rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t)1[\mathcal{A}^c(n)]\,dt + \mathbb{E}\sum_{n=\lfloor x^{1/3}\rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t)1[S(t;J) > x-1]1[\mathcal{A}(n)]\,dt.$$

Now, replacing (36) in the first expression of the preceeding inequality, computing the integral and using the exponential bound on $\mathbb{P}[\mathcal{A}^c(n)]$ lead to

$$\mathbb{E}\sum_{n=\lfloor x^{1/3}\rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t)1[\mathcal{A}^c(n)]\,dt \leq \sum_{n=\lfloor x^{1/3}\rfloor}^{\infty} \mathbb{P}[\mathcal{A}^c(n)] = o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x \to \infty.$$

Hence, applying the preceding estimate and conditioning on the length of T_0^r result in

$$\begin{aligned} I_3(x) &= \mathbb{E}\sum_{n=\lfloor x^{1/3}\rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t)1[S(t;J) > x-1, T_0^r > hx^\alpha, \mathcal{A}(n)]\,dt \\ &\quad + \mathbb{E}\sum_{n=\lfloor x^{1/3}\rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t)1[S(t;J) > x-1, T_0^r \leq hx^\alpha, \mathcal{A}(n)]\,dt + o\left(\frac{1}{x^{\alpha-1}}\right) \\ &\triangleq I_{31}(x) + I_{32}(x) + o\left(\frac{1}{x^{\alpha-1}}\right). \end{aligned} \tag{45}$$

Note that for all $\omega \in \mathcal{A}(n)$

$$\sum_{k=1}^{M} \tau_k(T_{0,n}^r) q_i^{(k)} \geq (1-\epsilon) n \sum_{k=1}^{M} q_i^{(k)} \nu_k \mu_k = (1-\epsilon) n\mu \sum_{k=1}^{M} q_i^{(k)} \pi_k = (1-\epsilon) n\mu q_i,$$

and therefore, for $t \in (T_n^r, T_{n+1}^r]$

$$\hat{f}(t) 1[\mathcal{A}(n)] \leq \sum_{i=1}^{\infty} q_i^{(J_{T_0^r})} q_i^{(J_{T_{n+1}^r})} e^{-q_i^{(J_{T_0^r})} T_0^r} e^{-\mu n q_i (1-\epsilon)} e^{-q_i^{(J_{T_{n+1}^r})} (t - T_n^r)} 1[\mathcal{A}(n)]. \tag{46}$$

Therefore, by using the preceeding inequality in $I_{31}(x)$, then completing the integration and applying $1 - e^{-x} \leq x$, $x \geq 0$, we derive

$$\begin{aligned}
I_{31}(x) &\leq \mathbb{E} \sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t) 1[T_0^r > hx^\alpha, \mathcal{A}(n)] \, dt \\
&\leq \left(\mathbb{E} \sum_{i=1}^{\infty} q_i^{(J_0)} e^{-q_i^{(J_0)} hx^\alpha} q_i^{(J_{T_{n+1}^r})} 1[T_0^r > hx^\alpha] (T_{n+1}^r - T_n^r) \right) \sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} e^{-\mu n q_i (1-\epsilon)} \\
&\leq H \mathbb{E} \left[1[T_0^r > hx^\alpha] (T_{n+1}^r - T_n^r) \sum_{i=1}^{\infty} q_i^{(J_0)} e^{-q_i^{(J_0)} hx^\alpha} \right],
\end{aligned}$$

where the last inequality uses $\sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} e^{-\mu n q_i (1-\epsilon)} = O\left(1/\bar{q}_i\right)$ and $q_i^{(J_{T_n^r})} \leq \bar{q}_i$. Hence, upper bounding the preceding sum and $\mathbb{E}[(T_{n+1}^r - T_n^r) 1[T_0^r > hx^\alpha]]$ similarly as in (13) and (38), respectively, we easily arrive at

$$I_{31}(x) = o\left(\frac{1}{x^{\alpha-1}} \right) \qquad \text{as} \qquad x \to \infty. \tag{47}$$

Next, in evaluating $I_{32}(x)$, we condition on the length of $T_{n+1}^r - T_n^r$

$$\begin{aligned}
I_{32}(x) = &\sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} \mathbb{E} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t) 1[S(t;J) > x-1, T_0^r \leq hx^\alpha, T_{n+1}^r - T_n^r > hx^\alpha, \mathcal{A}(n)] \, dt \\
&+ \sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} \mathbb{E} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t) 1[S(t;J) > x-1, T_0^r \leq hx^\alpha, T_{n+1}^r - T_n^r \leq hx^\alpha, \mathcal{A}(n)] \, dt. \qquad (48)
\end{aligned}$$

(49)

Thus, using (46), $q_i^{(J_0)} \leq \bar{q}_i$ and $1 - e^{-x} \leq x$ for $x \geq 0$, we upper bound the first term of the preceding equality

$$\begin{aligned}
&\mathbb{E} \sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t) 1[S(t;J) > x-1, T_0^r \leq hx^\alpha, T_{n+1}^r - T_n^r > hx^\alpha, \mathcal{A}(n)]\, dt \\
&\leq \mathbb{E} \sum_{i=1}^{\infty} \sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} \bar{q}_i e^{-(1-\epsilon)\mu n q_i} (1 - e^{-q_i^{(J_{T_{n+1}^r})}(T_{n+1}^r - T_n^r)}) 1[T_{n+1}^r - T_n^r > hx^\alpha] \\
&\leq H \mathbb{E}[(T_1^r - T_0^r) 1[T_1^r - T_0^r > hx^\alpha]] \sum_{i=1}^{\infty} \sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} (\bar{q}_i)^2 e^{-(1-\epsilon)\mu n q_i} \\
&= o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x \to \infty; \qquad (50)
\end{aligned}$$

the last estimate follows from $\mathbb{E}(T_1^r - T_0^r) < \infty$ and

$$\sum_{i=1}^{\infty} \sum_{n=\lfloor x^{\frac{1}{3}} \rfloor}^{\infty} (\bar{q}_i)^2 e^{-(1-\epsilon)\mu n q_i} \leq \sum_{i=1}^{\infty} \frac{(\bar{q}_i)^2}{q_i(1-\epsilon)\mu} = \frac{1}{(\min_{1 \leq k \leq M} \pi_k)^2 (1-\epsilon)\mu} < \infty.$$

Therefore, replacing expression (50) in the first sum of (48) yields as $x \to \infty$

$$\begin{aligned}
I_{32}(x) &= \mathbb{E} \sum_{n=\lfloor x^{1/3} \rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} \hat{f}(t) 1[S(T_{n+1}^r;J) > x, T_0^r \leq hx^\alpha, T_{n+1}^r - T_n^r \leq hx^\alpha, \mathcal{A}(n)]\, dt + o\left(\frac{1}{x^{\alpha-1}}\right) \\
&\leq \mathbb{E} \sum_{n=\lfloor x^{1/3} \rfloor}^{\lfloor \epsilon x^\alpha \rfloor} \int_{T_n^r}^{T_{n+1}^r} + \mathbb{E} \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} \int_{T_n^r}^{T_{n+1}^r} + \mathbb{E} \sum_{n=\lfloor g_\epsilon x^\alpha \rfloor}^{\infty} \int_{T_n^r}^{T_{n+1}^r} + o\left(\frac{1}{x^{\alpha-1}}\right) \\
&\triangleq I_{32}^{(1)}(x) + I_{32}^{(2)}(x) + I_{32}^{(3)}(x) + o\left(\frac{1}{x^{\alpha-1}}\right), \qquad (51)
\end{aligned}$$

for some $0 < \epsilon < g_\epsilon$. Now, recalling the definition of $S(u,t;J)$ and using similar arguments as in (40) and (41), it is easy to show

$$\begin{aligned}
&1[S(T_{n+1}^r;J) > x-1, T_0^r \leq hx^\alpha, T_{n+1}^r - T_n^r \leq hx^\alpha] \leq \\
&1[S(T_0^r, T_n^r;J) > (1-2\epsilon)x] + 1[S(hx^\alpha;J) > \epsilon x - 1/2] + 1[S(T_n^r, T_n^r + hx^\alpha;J) > \epsilon x - 1/2], \qquad (52)
\end{aligned}$$

and, therefore, replacing (46) in $I_{32}^{(1)}(x)$ and completing the integration result in

$$\begin{aligned}
I_{32}^{(1)}(x) \leq \mathbb{E} &\sum_{n=\lfloor x^{1/3} \rfloor}^{\lfloor \epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} q_i^{(J_0)} e^{-(1-\epsilon) n q_i \mu} (1 - e^{-q_i^{(J_{T_{n+1}^r})}(T_{n+1}^r - T_n^r)}) 1[S(T_0^r, T_n^r;J) > x(1-2\epsilon)] \\
&+ 2\epsilon x^\alpha \mathbb{P}[\bar{S}(hx^\alpha) > \epsilon x - 1/2],
\end{aligned}$$

where h is small enough to ensure $\mathbb{E}\bar{S}(hx^\alpha) \leq \epsilon x/2$ for large x. Next, applying Lemma 4, $q_i^{(J_0)} \leq H q_i$, $1 - e^{-x} \leq x$ ($x \geq 0$), the fact that $(T_{n+1}^r - T_n^r)$ and $1[S(T_0^r, T_n^r;J) > (1-2\epsilon)x]$ are conditionally

independent given $J_{T_n^r}$ and $\mathbb{E}[T_1^r - T_0^r] < \infty$ renders, as $x \to \infty$,

$$I_{32}^{(1)}(x) \leq H \sum_{n=\lfloor x^{1/3} \rfloor}^{\lfloor \epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)n\mu q_i} \mathbb{E}[\mathbb{E}[T_{n+1}^r - T_n^r | J_{T_n^r}] \mathbb{P}[S(T_0^r, T_n^r; J) > x(1-2\epsilon) | J_{T_n^r}]] + o\left(\frac{1}{x^{\alpha-1}}\right)$$
$$\leq H \sum_{n=\lfloor x^{1/3} \rfloor}^{\lfloor \epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)n\mu q_i} \mathbb{P}[\bar{S}(T_n^r - T_0^r) > x(1-2\epsilon)] + o\left(\frac{1}{x^{\alpha-1}}\right),$$

Now, by using the argument as in inequality (42) and Lemmas 2 and 7, we derive as $x \to \infty$

$$I_{32}^{(1)}(x) \leq H \sum_{n=\lfloor x^{\frac{1}{3}} \rfloor}^{\lfloor \epsilon x^\alpha \rfloor} \frac{1}{n^{2-\frac{1}{\alpha}}} n\mathbb{P}[T_1^r - T_0^r > hx^\alpha] + o\left(\frac{1}{x^{\alpha-1}}\right) = o\left(\frac{1}{x^{\alpha-1}}\right), \tag{53}$$

where ϵ is smaller than h in Lemma 7. Now, we estimate $I_{32}^{(2)}(x)$ by replacing (46) in $I_{32}^{(2)}(x)$, completing the integration and applying similar arguments as in (52) and, therefore, as $x \to \infty$

$$I_{32}^{(2)}(x) \leq \mathbb{E} \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} q_i^{(J_0)} e^{-(1-\epsilon)nq_i\mu} (1 - e^{-q_i^{(J_{T_{n+1}^r})}(T_{n+1}^r - T_n^r)}) 1[S(T_0^r, T_n^r; J) > (1-2\epsilon)x] + o\left(\frac{1}{x^{\alpha-1}}\right).$$

Then, by using $\max(q_i^{(J_0)}, q_i^{(J_{T_{n+1}^r})}) \leq Hq_i$, $1 - e^{-x} \leq x$ ($x \geq 0$), $\mathbb{E}[T_1^r - T_0^r] < \infty$, the fact that $(T_{n+1}^r - T_n^r)$ and $1[S(T_0^r, T_n^r; J) > (1-2\epsilon)x]$ are conditionally independent given $J_{T_n^r}$, we obtain as $x \to \infty$

$$I_{32}^{(2)}(x) \leq H \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)nq_i\mu} \mathbb{E}\left[\mathbb{E}[T_{n+1}^r - T_n^r | J_{T_n^r}] \mathbb{P}\left[S\left(T_0^r, T_n^r; J\right) > x(1-2\epsilon) \,\middle|\, J_{T_n^r}\right]\right]$$
$$+ o\left(\frac{1}{x^{\alpha-1}}\right)$$
$$\leq H \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)nq_i\mu} \mathbb{P}\left[S\left(T_0^r, T_{\lfloor g_\epsilon x^\alpha \rfloor}^r; J\right) > x(1-2\epsilon)\right] + o\left(\frac{1}{x^{\alpha-1}}\right),$$

Define

$$\mathcal{B}(x) \triangleq \cap_{1 \leq k \leq M} \{\tau_k(T_{0,\lfloor g_\epsilon x^\alpha \rfloor}^r) \leq (1+\epsilon)\mu_k \nu_k g_\epsilon x^\alpha\}.$$

Then, as $x \to \infty$

$$I_{32}^{(2)}(x) \leq H \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)\mu n q_i} \mathbb{P}\left[S\left(T_0^r, T_{\lfloor g_\epsilon x^\alpha \rfloor}^r; J\right) > x(1-2\epsilon), \mathcal{B}(x)\right]$$
$$+ H \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)\mu n q_i} \mathbb{P}[\mathcal{B}^c(x)] + o\left(\frac{1}{x^{\alpha-1}}\right). \tag{54}$$

Due to the Weak Law of Large Numbers, $\mathbb{P}[\tau_k(T^r_{0,g_\epsilon x^\alpha}) > (1+\epsilon)\mu_k\nu_k g_\epsilon x^\alpha] \to 0$, implying $\mathbb{P}[\mathcal{B}^c(x)] \to 0$ as $x \to \infty$, which, in conjunction with Lemma 2, yields as $x \to \infty$

$$H \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)\mu n q_i} \mathbb{P}[\mathcal{B}^c(x)] = o(1) \sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} n^{-2+\frac{1}{\alpha}} = o\left(\frac{1}{x^{\alpha-1}}\right). \tag{55}$$

Next, we estimate the probability $\mathbb{P}[S(T_0^r, T_{\lfloor g_\epsilon x^\alpha \rfloor}; J) > (1-2\epsilon)x, \mathcal{B}(x)]$. Let $B_i(u,t;J)$, $0 < u < t$, be a random variable indicating whether item i is requested in $[-t,-u)$. Define $S^*(x) = \sum_{i=1}^{\infty} B_i^*(x)$, where $\{B_i^*(x), i \geq 1\}$ is a sequence of independent Bernoulli random variables with $\mathbb{P}[B_i^*(x) = 1] = 1 - e^{-(1+\epsilon)\sum_{k=1}^M q_i^{(k)} \mu \pi_k g_\epsilon x^\alpha}$; similarly as before, $S^*(x)$ is constructed non-decreasing in x. Then, for every $\omega \in \mathcal{B}(x)$

$$\begin{aligned}\mathbb{P}_{\sigma_{T^r_{\lfloor g_\epsilon x^\alpha \rfloor}}}\left[B_i\left(T_0^r, T^r_{\lfloor g_\epsilon x^\alpha \rfloor}; J\right) = 1\right] &= 1 - e^{-\sum_{k=1}^M q_i^{(k)} \tau_k(T^r_{0,\lfloor g_\epsilon x^\alpha \rfloor})} \\ &\leq \mathbb{P}[B_i^*(x) = 1].\end{aligned}$$

Therefore, by stochastic dominance, for every $\omega \in \mathcal{B}(x)$

$$\mathbb{P}_{\sigma_{T^r_{\lfloor g_\epsilon x^\alpha \rfloor}}}\left[S\left(T_0^r, T^r_{\lfloor g_\epsilon x^\alpha \rfloor}; J\right) > (1-2\epsilon)x\right] \leq \mathbb{P}\left[S^*(x) > (1-2\epsilon)x\right]. \tag{56}$$

If we select $g_\epsilon = \frac{1-3\epsilon}{(1+\epsilon)c\mu\Gamma\left[1-\frac{1}{\alpha}\right]^\alpha}$, it is easy to check, using Lemma 3, that for all x large enough

$$\mathbb{E}S^*(x) = \sum_{i=1}^{\infty} (1 - e^{-(1+\epsilon)\sum_{k=1}^M \pi_k \mu g_\epsilon x^\alpha q_i^{(k)}}) < (1-3\epsilon)x.$$

This inequality, (56) and Lemma 4 imply, for all x large enough,

$$\mathbb{P}\left[S\left(T_0^r, T_{\lfloor g_\epsilon x^\alpha \rfloor}; J\right) > (1-2\epsilon)x, \mathcal{B}(x)\right] \leq \mathbb{P}\left[S^*(x) > (1-2\epsilon)x\right] \leq He^{-\theta_\epsilon x},$$

for some positive constant θ_ϵ. Therefore, by using Lemma 2, the upper bound the first expression in (54) is

$$\sum_{n=\lfloor \epsilon x^\alpha \rfloor}^{\lfloor g_\epsilon x^\alpha \rfloor} \sum_{i=1}^{\infty} (q_i)^2 e^{-(1-\epsilon)\mu n q_i} \mathbb{P}\left[S^*(x) > (1-2\epsilon)x\right] = o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x \to \infty,$$

which in conjunction with (55) implies

$$I_{32}^{(2)}(x) = o\left(\frac{1}{x^{\alpha-1}}\right) \qquad \text{as} \qquad x \to \infty. \tag{57}$$

Finally, after replacing (46) in $I_{32}^{(3)}(x)$, computing the integral, and applying first $1 - e^{-x} \leq x$, $x \geq 0$ and then (1), we obtain

$$I_{32}^{(3)}(x) \leq \sum_{\lfloor g_\epsilon x^\alpha \rfloor}^{\infty} \sum_{i=1}^{\infty} (q_i)^2 (1+\epsilon)\mu e^{-(1-\epsilon)\mu n q_i}.$$

Now, by using Lemma 2 and replacing the expression for g_ϵ, we compute

$$\limsup_{x\to\infty} \frac{I_{32}^{(3)}(x)}{\mathbb{P}[R > x]} \leq K(\alpha) \frac{(1+\epsilon)^{2-\frac{1}{\alpha}}(1-\epsilon)^{\frac{1}{\alpha}}}{(1-3\epsilon)^{1-\frac{1}{\alpha}}}.$$

By passing $\epsilon \to 0$ in the last inequality and then replacing it together with estimates (57) and (53) in (51), we derive

$$\limsup_{x\to\infty} \frac{I_{32}(x)}{\mathbb{P}[R > x]} \leq K(\alpha).$$

Finally, (51), (47), (45), (35), (43) and Proposition 2 conclude the proof.

$\diamond$

8 Appendix

Proof of Proposition 1: By Theorem 1, for any finite N, the stationary search cost is given by

$$\mathbb{P}[C^N > x] = \mathbb{E}\int_0^\infty \sum_{i=1}^N q_{i,N}^{(J_0)} q_{i,N}^{(J_{-t})} e^{-\hat{q}_{i,N}t} \mathbb{P}[S_{i,N}(t;J) > x-1]\, dt. \tag{58}$$

Clearly, the term under the integral in the preceding equation converges to the corresponding term in (10) as $N \to \infty$. Hence, in order to apply the Dominated Convergence Theorem, it remains to show that uniformly in N, the integrand in (58) is bounded by an integrable function. To this end, let $\hat{q}_{i,N}$, $i \geq 1$ correspond to empirical distribution, defined in (8) with $q_i^{(k)}$ replaced by $q_{i,N}^{(k)}$. Then, since $\frac{1}{\sum_{i=1}^N q_i^{(k)}} \searrow 1$ as $N \to \infty$, there exists $N_0 \geq 1$, such that for all $N \geq N_0$, $1 \leq i \leq N$ and $1 \leq k \leq M$

$$q_i^{(k)} \leq q_{i,N}^{(k)} \leq 2q_i^{(k)}.$$

Thus, the function under the integral in (58) is almost surely bounded by

$$4\sum_{i=1}^\infty q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t}. \tag{59}$$

Since $-d(e^{-\hat{q}_i t}) = e^{-\hat{q}_i t} d(\sum_{k=1}^M q_i^{(k)} \int_{-t}^0 1[J_u = k] du) = e^{-\hat{q}_i t} q_i^{(J_{-t})} dt$, and, due to ergodicity, $\hat{q}_i t = \sum_{k=1}^M q_i^{(k)} \int_{-t}^0 1[J_u = k]\, du \to \infty$ as $t \to \infty$ a.s., we conclude that the function in (59) is integrable, i.e.

$$\mathbb{E}\int_0^\infty 4\sum_{i=1}^\infty q_i^{(J_0)} q_i^{(J_{-t})} e^{-\hat{q}_i t}\, dt = -4\mathbb{E}\int_0^\infty \sum_{i=1}^\infty q_i^{(J_0)} d(e^{-\hat{q}_i t}) = 4.$$

$\diamond$

Proof of Lemma 4: Let $m_i \triangleq \mathbb{E}B_i$, $i \geq 1$. For an arbitrary $0 < \epsilon < 1$

$$\mathbb{P}[|S - m| > m\epsilon] = \mathbb{P}[S > m(1+\epsilon)] + \mathbb{P}[-S > -m(1-\epsilon)]. \tag{60}$$

Now, using Markov inequality, for any $\theta > 0$ we obtain

$$\mathbb{P}[S > m(1+\epsilon)] = \mathbb{P}[e^{\theta S} > e^{\theta m(1+\epsilon)}] \leq \frac{\mathbb{E}e^{\theta S}}{e^{\theta m(1+\epsilon)}}. \tag{61}$$

Since, $\{B_i, i \geq 1\}$ are independent Bernoulli random variables

$$\mathbb{E}e^{\theta S} = \prod_{i=1}^\infty \mathbb{E}e^{\theta B_i} = \prod_{i=1}^\infty (e^\theta m_i + (1-m_i)) \leq \prod_{i=1}^\infty e^{m_i(e^\theta - 1)},$$

and, therefore, using (61), we derive

$$\mathbb{P}[S > m(1+\epsilon)] \leq e^{m(e^{\theta}-1-\theta(1+\epsilon))}.$$

Thus, we can choose $\theta > 0$ such that $e^{\theta} - 1 - \theta(1+\epsilon) = -\theta_{\epsilon}^{(1)} < 0$. Similarly, we get that the second expression of (60) is bounded by $e^{-\theta_{\epsilon}^{(2)} m}$ for some $\theta_{\epsilon}^{(2)} > 0$. By taking $\theta_{\epsilon} = \min(\theta_{\epsilon}^{(1)}, \theta_{\epsilon}^{(2)})$, we complete the proof.

$\diamond$

References

[1] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliviera. Characterizing reference locality in the WWW. In *Proceedings of the Fourth International Conference on Parallel and Distributed Information Systems*, 1996.

[2] F. Baccelli and P. Bremaud. *Elements of Queueing Theory: Palm-Martingale Calculus and Stochastic Recurence*. Springer Verlag, 1994.

[3] J. L. Bentley and C. C. McGeoch. Amortized analysis of self-organizing sequential search heuristics. *Communications of the ACM*, 28(4):404–411, 1985.

[4] J. R. Bitner. Heuristics that dynamically organize data structures. *SIAM J. Comput.*, 8:82–110, 1979.

[5] W. A. Borodin, S. Irani, P. Raghavan, and B. Schieber. Competitive paging with locality of reference. *Journal of Computer and System Science*, 50(2):244–258, 1995.

[6] P. J. Burville and J. F. C. Kingman. On a model for storage and search. *Journal of Applied Probability*, 10:697–701, 1973.

[7] E. Cinlar. *Introduction to Stochastic Processes*. Prentice-Hall, 1975.

[8] E. G. Coffman and P. R. Jelenković. Performance of the Move-To-Front Algorithm with Markov-Modulated Request Sequences. *Operations Research Letters*, 25:109–118, 1999.

[9] R. P. Dobrow and J. A. Fill. The move-to-front rule for self-organizing lists with Markov dependent requests. In D. Aldous, P. Diaconis, J. Spencer, and J. M. Steele, editors, *Discrete Probability and Algorithms*, pages 57–80. Springer-Verlag, 1995.

[10] J. A. Fill. An exact formula for the move-to-front rule for self-organizing lists. *Journal of Theoretical Probability*, 9(1):113–159, 1996.

[11] J. A. Fill. Limits and rate of convergence for the distribution of search cost under the move-to-front rule. *Theoretical Computer Science*, 164:185–206, 1996.

[12] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collector, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39:207–229, 1992.

[13] S. Irani, A. R. Karlin, and S. Philips. Strongly competitive algorithms for paging with locality of reference. *SIAM J. Comput.*, 25(3):477–497, June 1996.

[14] P. R. Jelenković. A new asymptotic analysis of the move-to-front search cost distribution. In *4th International Seminar on Avarage-Case Analysis of Algorithms*, Princeton, NJ, July 1998.

[15] P. R. Jelenković. Asymptotic approximation of the move-to-front search cost distribution and least-recently-used caching fault probabilities. *Annals of Applied Probability*, 9(2):430–464, 1999.

[16] P. R. Jelenković and A. A. Lazar. Subexponential asymptotics of a Markov-modulated random walk with queueing applications. *Journal of Applied Probability*, 35(2):325–347, June 1998.

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

[18] J. McCabe. On serial files with relocatable records. *Operations Research*, 13:609–618, 1965.

[19] S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 1979.

[20] R. M. Phataford. On the transition probabilities of the move-to-front scheme. *Journal of Applied Probability*, 31:570–574, 1994.

[21] R. Rivest. On self-organizing sequential search heuristics. *Communications of the ACM*, 19(2):63–67, 1976.

[22] E. R. Rodrigues. The performance of the move-to-front scheme under some particular forms of Markov requests. *Journal of Applied Probability*, 32(4):1089–1102, 1995.

[23] D. D. Sleator and R. E. Tarjan. Amortized efficiency of list update and paging rules. *Communications of the ACM*, 28(2):202–208, 1985.