The Persistent-Access-Caching Algorithm

Predrag R. 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

April 2004

Abstract

Caching is widely recognized as an effective mechanism for improving the performance of the World Wide Web. One of the key components in engineering the Web caching systems is designing document placement/replacement algorithms for updating the collection of cached documents. The main objectives of such a policy are the high cache hit ratio, ease of implementation, low complexity and adaptability to the fluctuations in access patterns. These objectives are satisfied by the widely used heuristic is the Least-Recently-Used (LRU) cache replacement rule. However, in the context of the independent reference model, the LRU policy can significantly underperform the optimal Least-Frequently-Used (LFU) algorithm. On the other hand, the benefit of the LFU rule is counterbalanced by its higher implementation complexity and lower adaptability to changes in access frequencies.

To alleviate this problem, we introduce a new LRU-based rule, termed the Persistent-Access-Caching (PAC), which essentially preserves all of the desirable attributes of the LRU scheme. Furthermore, for the independent reference model and generalized Zipf's law request probabilities, we prove that, for large cache sizes, the PAC policy performs arbitrarily close to the optimal LFU algorithm. The PAC algorithm makes the replacement decisions based on the references collected during the preceding interval of fixed length that, for large caches, adds a negligible additional complexity when compared to the ordinary LRU policy.

Keywords: persistent-access-caching, least-recently-used caching, least-frequently-used caching, move-to-front searching, generalized Zipf's law distributions, heavy-tailed distributions, Web caching, cache fault probability, average-case analysis

†Supported by the NSF Grant No. 0092113.
‡Supported by IBM PhD Fellowship.
Technical Report EE2004-03-05, Department of Electrical Engineering, Columbia University, April 2004. Submitted for publication.

1 Introduction

Since the recent invention of the World Wide Web (WWW), there have been an explosive growth in multimedia information content and services that include data, audio, video, software downloads, remote service hosting, etc. These distributed multimedia content and services are now integral part of modern communication networks (e.g., the Internet) and, therefore, are redefining the role of networking to incorporate the storage and service of information, in addition to the traditional task of information transfer. Since network information and its access are massively distributed, and the same content is repeatedly used by groups of users, it is clear that bringing some of the more popular items closer to the end-users can improve the network performance, e.g., reduce the download latency and network congestion. This type of information replication and redistribution system is often termed Web caching.

One of the key components of engineering efficient Web caching systems is designing document placement/replacement algorithms that are selecting and possibly dynamically updating a collection of frequently accessed documents. The design of these algorithms has to be done with special care since the latency and network congestion may actually increase if documents with low access frequency are cached. Thus, the main objective is to achieve high cache hit ratios, while maintaining ease of implementation and scalability. Furthermore, these algorithms need to be self-organizing and robust since the document access patterns exhibit a high degree of spatial as well as time fluctuations. The well-known heuristic named the Least-Recently-Used (LRU) cache replacement rule satisfies all of the previously mentioned attributes and, therefore, represents a basis for designing many practical replacement algorithms. However, as shown in [8] in the context of the stationary independent reference model with generalized Zipf's law requests, this rule is by a constant factor away from the optimal frequency algorithm that keeps in the cache most frequently used documents, i.e., replaces Least-Frequently-Used (LFU) items. On the other hand, the drawback of the LFU algorithm is that it needs to know (measure) the document access frequencies and employ aging schemes based on reference counters in order to cope with evolving access patterns, which results in high complexity. In the context of database disk buffering, [14] proposes a modification of the LRU policy, called LRU-K, that uses the information of the last K reference times for each document in order to make replacement decisions. It is shown in [14] that the fault probability of the LRU-K policy approaches, as K increases, the performance of the optimal LFU scheme. However, practical implementation of the LRU-K policy would still be the same order of complexity as the LFU rule. Furthermore, for larger values of K, that might be required for nearly optimal performance, the adaptability of this algorithm to changes in traffic patterns will be significantly reduced.

In this paper we design a new LRU-based policy, termed the Persistent-Access-Caching (PAC) rule, that essentially preserves all the desirable features of LRU caching, while achieving arbitrarily close performance to the optimal LFU algorithm. Furthermore, the PAC algorithm has only negotiable additional complexity in comparison to the widely used LRU policy. We analyze the PAC replacement scheme with i.i.d. requests that arrive at Poisson time points. If a requested document at time t, say i, is not found in the cache, it is placed inside only if it is requested more than $k-1$ times in interval $(t-\beta, t]$. The parameters k and β are the fixed design values of the PAC algorithm, whose detailed description will be provided in the following section. When the frequency of requesting a page i is equal to the generalized Zipf's law c/i^{α}, $\alpha > 0$, we prove that the cache fault probability asymptotically approaches, as k increases, the performance of the optimal LFU algorithm. It is surprising that even for the small values of k, the performance ratio between the PAC and optimal algorithms significantly improves when compared to the ordinary LRU; for example, in the case of $\alpha > 1$, this ratio drops from approximately 1.78 for $k = 1$ to $1.18, 1.08$ for $k = 2, 3$, respectively. Furthermore, we show that the derived asymptotic results and simulation experiments match each other very well, even for relatively small cache sizes.

Our analytical approach uses probabilistic (average-case) analysis that exploits the novel large deviation technique and asymptotic results that were recently developed in [8, 11]. The computation of the LRU fault probability is mathematically equivalent to the evaluation of the search cost distribution for the related Move-To-Front (MTF) searching scheme. For recent work on average case analysis of MTF and LRU algorithms see [6, 5, 8, 11] and the references therein. For an alternative combinatorial (competative) approach to analyzing LRU caches the reader can consult [2, 13] and the references therein.

This paper is organized as follows. In Section 2, first, we formally describe the PAC policy with a Poisson reference model. Then, using the Poisson decomposition/superposition properties, we develop a representation theorem for the stationary search cost of the related, Persistent Move-To-Front algorithm. This representation formula, in conjunction with the results on Poisson processes derived in Subsection 2.1, provide the starting point for proving our main theorems in Section 3. Informally, our main results show that for large cache sizes, independent reference model, and generalized Zipf's law request distributions, the fault probability of the PAC algorithm approaches the optimal LFU policy with a negligible additional complexity. Furthermore, in Section 4, extensive numerical experiments show an excellent agreement between our analytical results and simulations. The paper is concluded in Section 5 with a brief discussion of our results and their possible extensions.

2 Model description and preliminary results

Consider a set $L = \{1, 2, \dots, N\}$ of N documents (possibly infinite), out of which x can be stored in an easily accessible location, called cache. The remaining $N - x$ documents (items) are placed outside of the cache in a slower access medium. Documents are requested at moments $\{\tau_n\}_{n\geq 1}$ that represent a positive increasing sequence of Poisson points of unit rate. Furthermore, define a sequence of i.i.d. random variables $\{R_n\}_{n\geq 1}$, independent from $\{\tau_n\}_{n\geq 1}$, where $\{R_n = i\}$ represents a request for item i at time τ_n. We denote request probabilities as $\mathbb{P}[R_n = i] = q_i$ and, without loss of generality, we assume $q_1 \geq q_2 \geq \dots$; let $M^{(q_i)}(u, t)$ be the number of requests for item i in an open interval (u, t). Documents stored in the cache are ordered in a list, which is sequentially searched upon a request for a document and is updated as follows. If a requested document at the moment τ_n, say i, is found in the cache, we have a cache hit. In this case, if $M^{(q_i)}(\tau_n - \beta, \tau_n) \geq k - 1$, item i is moved to the front of the list while documents that were in front of item i are shifted one position down; otherwise, the list stays unchanged. Furthermore, if document i is not found in the cache, we call it a cache miss or fault. Then, similarly as before, if $M^{(q_i)}(\tau_n - \beta, \tau_n) \geq k - 1$, document i is brought to the first position of the cache list and the least recently moved item, i.e. the one at the last position of the list, is evicted from the cache. Previously described cache replacement policy we call PAC(β, k) algorithm. The performance measure of interest is the cache fault probability, i.e. the probability that a requested document is not found in the cache.

Analyzing the PAC(β, k) algorithm is equivalent to investigating the corresponding Move-To-Front (MTF) scheme that is defined as follows. Consider a list $L = \{1, 2, \dots, N\}$ and a process of requests for documents determined by $\{R_n\}_{n\geq 1}$ and $\{\tau_n\}_{n\geq 1}$ as in the preceding paragraph. When a request for a document arrives, say $R_n = i$, the list is searched and the requested item is moved to the front of the list only if $M^{(q_i)}(\tau_n - \beta, \tau_n) \geq k - 1$; otherwise the list stays unchanged. Previously described searching algorithm we call Persistent-MTF, PMTF(β, k). The performance measure of interest for this algorithm is the search cost $C_n^{(N)}$ that represents the position of the requested document at time τ_n.

Now, we claim that computing the cache fault probability of the PAC(β, k) algorithm is equivalent to evaluating the tail of the searching cost $C_n^{(N)}$ of the PMTF(β, k) searching scheme. Note that the fault probability of the PAC(β, k) algorithm stays the same regardless of the ordering of documents in the slower access medium. In particular, these documents can be also ordered in the increasing order of the last times they are moved to the front of the cache list. Therefore, it is intuitively clear that the fault probability of the PAC(β, k) policy for the cache of size x after the nth request is the same as the probability that the search cost of the PMTF(β, k) algorithm is greater than x, i.e. $\mathbb{P}[C_n^{(N)} > x]$. Hence, even though PAC(β, k) and PMTF(β, k) belong to different application areas, their performance analysis is essentially equivalent. Thus, in the rest of the paper we investigate the tail of the stationary search cost distribution.

First, we prove the convergence of the search cost $C_n^{(N)}$ to stationarity. Suppose that the system starts at $t = 0$ with initial conditions given by an arbitrary initial permutation Π_0 of the list and a sequence of requests $\mathcal{R}_0 = \{(\tau_{0i}, R_{0i})\}_{i\geq 1}$ in interval $(-\beta, 0)$; $\tau_{0i} \in (-\beta, 0)$ is the time of the ith initial request R_{0i}. Denote the sequence of time points of requests for document i as $\{\tau_n^{(q_i)}\}_{n\geq 1}$. Then, by the assumptions on the request process $\{R_n\}$, arrival times $\{\tau_n\}$ and Poisson decomposition theorem, we conclude that processes $\{\tau_n^{(q_i)}\}_{n\geq 1}$, $i \geq 1$, are Poisson and independent.

In order to prove the convergence of $C_n^{(N)}$ to stationarity, we define another process of Poisson points of unit rate on the negative part of the real line $\{\tau_{-n}, n \geq 0\}$ and set $\tau_0 = 0$. Also, we define a sequence of i.i.d. random variables, $\{R_{-n}\}_{n\geq 0}$, independent from $\{\tau_{-n}\}_{n\geq 0}$, where $\mathbb{P}[R_{-n} = i] = q_i$. Now, for each n we construct a PMTF(β, k) algorithm starting at τ_{-n}, with a sequence of requests $\{R_{-m} : m = 0, 1, \ldots, n-1\}$ at times $\{\tau_{-m} : m = 0, 1, \ldots, n-1\}$ and having the same initial condition as in the previous paragraph, given by Π_0 and $\mathcal{R}_0$ in interval $(\tau_{-n} - \beta, \tau_{-n})$; let $C_{-n}^{(N)}$ be the search cost at τ_0. Note that in this construction we assume that for the PMTF(β, k) algorithm starting at τ_{-n} there is no request at time τ_{-n}. Now, if we consider the shift mapping $R_{n-k} \to R_{-k}$ and $\tau_{n-k} \to \tau_{-k}$ for $k = 0, 1, \ldots n-1$, we conclude that, since the corresponding sequences are equal in distribution, the search costs $C_{-n}^{(N)}$ and $C_n^{(N)}$ are also equal in distribution, i.e. $C_n^{(N)} \stackrel{d}{=} C_{-n}^{(N)}$. Thus, instead of computing the tail of the search cost $C_n^{(N)}$, we continue with evaluating the tail of $C_{-n}^{(N)}$. In this regard, we define a sequence of stopping times $\{T_i^{(-n)}\}_{n\geq 1}$, where $-T_i^{(-n)}$ represents the last time before $t = 0$ that item i was moved to the front of the list in the case of the PMTF(β, k) algorithm that started at τ_{-n}; if item i is not moved in $(\tau_{-n}, 0)$, we take that $T_i^{(-n)} = -\tau_{-n}$ a.s.. Next, we define stopping times T_i, $i \geq 1$, as

$$T_i \triangleq -\sup\{\tau_{-n}^{(q_i)} < 0 : \tau_{-n}^{(q_i)} - \tau_{-n-k+1}^{(q_i)} \leq \beta\}, \tag{1}$$

where process $\{\tau_{-n}^{(q_i)}\}_{n\geq 0}$ contains the moments of requests for document i; again, from the assumptions on the request process $\{R_{-n}\}_{n\geq 0}$ and arrival times $\{\tau_{-n}\}_{n\geq 0}$, by Poisson decomposition theorem, processes $\{\tau_{-n}^{(q_i)}\}_{n\geq 1}$, $i \geq 1$, are Poisson and mutually independent. Next, from the definitions of T_i and $T_i^{(-n)}$, we conclude that given the event $\{T_i^{(-n)} < -\tau_{-n} - \beta\}$, equality $T_i = T_i^{(-n)}$ a.s. holds. Therefore, the complementary sets of events are the same, i.e. $\{T_i \geq -\tau_{-n} - \beta\} = \{T_i^{(-n)} \geq -\tau_{-n} - \beta\}$. Then, given the previous observations, we bound the tail of the search cost $C_{-n}^{(N)}$ as

$$\begin{aligned}\mathbb{P}[C_{-n}^{(N)} > x, R_0 = i, T_i^{(-n)} < -\tau_{-n} - \beta] \leq \mathbb{P}[C_{-n}^{(N)} > x, R_0 = i] \leq \\ \mathbb{P}[C_{-n}^{(N)} > x, R_0 = i, T_i^{(-n)} < -\tau_{-n} - \beta] + \mathbb{P}[C_{-n}^{(N)} > x, R_0 = i, T_i^{(-n)} \geq -\tau_{-n} - \beta].\end{aligned} \tag{2}$$

Next, since the search cost $C_{-n}^{(N)}$ is equal to the number of different documents that are moved to the front of the list from the last time that item i was requested, we derive

$$\mathbb{P}[C_{-n}^{(N)} > x, R_0 = i, T_i^{(-n)} < -\tau_{-n} - \beta] = \mathbb{P}\left[R_0 = i, \sum_{j\neq i} 1[T_j^{(-n)} < T_i^{(-n)} < -\tau_{-n} - \beta] \geq x\right]$$

$$= q_i \mathbb{P}\left[\sum_{j\neq i} 1[T_j < T_i < -\tau_{-n} - \beta] \geq x\right],$$

where the last equality follows from the independence of processes $\{R_{-n}\}_{n\geq 1}$ and $\{\tau_{-n}\}_{n\geq 0}$ and the equality $T_i = T_i^{(-n)}$ a.s., $i \geq 1$, on $\{T_i < -\tau_{-n} - \beta\}$. Thus, since $-\tau_{-n} \to \infty$ a.s. as $n \to \infty$, we conclude, by monotone convergence theorem,

$$\lim_{n\to\infty} \sum_{i=1}^{N} \mathbb{P}[C_{-n}^{(N)} > x, R_0 = i, T_i < -\tau_{-n} - \beta] = \sum_{i=1}^{N} q_i \mathbb{P}\left[\sum_{j\neq i} 1[T_j < T_i] \geq x\right]. \tag{3}$$

Next, note that

$$\mathbb{P}[T_i \geq -\tau_{-n} - \beta] \leq \mathbb{P}[T_i > n(1-\epsilon)] + \mathbb{P}[-\tau_{-n} \leq n(1-\epsilon) + \beta]. \tag{4}$$

Then, due to the strong law of large numbers, since β is some finite constant,

$$\lim_{n\to\infty} \mathbb{P}[-\tau_{-n} \leq n(1-\epsilon) + \beta] = 0. \tag{5}$$

Furthermore, since $T_i < \infty$ a.s., we obtain

$$\lim_{n\to\infty} \mathbb{P}[T_i > n(1-\epsilon)] = 0. \tag{6}$$

Finally, equality of the events $\{T_i^{(-n)} \geq -\tau_{-n} - \beta\} = \{T_i \geq -\tau_{-n} - \beta\}$, independence of processes $\{R_{-n}\}_{n\geq 0}$ and $\{\tau_{-n}\}_{n\geq 0}$ and (4) imply

$$\lim_{n\to\infty} \sum_{i=1}^{N} \mathbb{P}[R_0 = i, T_i^{(-n)} \geq -\tau_{-n} - \beta] \leq \lim_{n\to\infty} \mathbb{P}[-\tau_{-n} \leq n(1-\epsilon)+\beta] + \lim_{n\to\infty} \sum_{i=1}^{N} q_i \mathbb{P}[T_i > n(1-\epsilon)+\beta] = 0,$$

where in the last equality we applied (5), (6) and the monotone convergence theorem. The previous expression, in conjunction with (3) and (2), implies the following result:

Lemma 1 *For any $1 \leq N \leq \infty$, arbitrary initial conditions $(\Pi_0, \mathcal{R}_0)$ and any $x \geq 0$, the search cost $C_n^{(N)}$ converges in distribution to $C^{(N)}$ as $n \to \infty$, where*

$$\mathbb{P}[C^{(N)} > x] \triangleq \sum_{i=1}^{N} q_i \mathbb{P}[S_i(T_i) \geq x]; \tag{7}$$

we define $S_i(t) \triangleq \sum_{j\neq i} 1[T_j < t]$, $i \geq 1$.

Remark 1 Note that the assumption on the Poisson request times is crucial for the analysis of the PAC algorithm. In particular, this assumption in conjunction with i.i.d. request process implies the independence of the stopping times T_i, $1 \leq i \leq N$. Otherwise, if the request times are not Poisson, e.g. discrete time arrivals, these variables may not be independent, which would make the analysis intractable. The Poisson embedding technique for i.i.d. requests was first introduced in [5].

In order to evaluate the tail of the stationary search cost $C^{(N)}$, we need estimates for random times T_i, $i \geq 1$. In the next section we prove both lower and upper bounds for T_i, which we use to prove our main results in section 3.

2.1 Preliminary results on Poisson processes

Let $\{\tau_n^{(q)}\}_{n\geq 1}$ be a positive increasing sequence of Poisson points with rate q and let $M^{(q)}(u,t)$ be the number of Poisson points in an open interval (u,t). Now, we investigate the distribution of the first time T such that the interval $[T, T+\beta)$ contains at least $k \geq 1$ Poisson points, i.e.,

$$T \triangleq \inf\{\tau_n^{(q)} : \tau_{n+k-1}^{(q)} - \tau_n^{(q)} \leq \beta\}. \tag{8}$$

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.

Lemma 2 *For any $\epsilon > 0$, there exists $q_0 > 0$, such that for all $0 < q \leq q_0$,*

$$\mathbb{P}[T > t] \leq e^{-\frac{q^k \beta^{k-1}}{(k-1)!}(1-\epsilon)t} + 2e^{-h\epsilon q^{k-1} t}. \tag{9}$$

Proof: For $k = 1$ the bound trivially holds since $T \equiv \nu_1$ and, thus, we assume that $k \geq 2$.

Let $\{\tau_n^{(d)}\}_{n\geq 1}$ be a process obtained from $\{\tau_n^{(q)}\}_{n\geq 1}$ by deleting some of the points according to the following rule. Starting at the first point $\tau_1^{(q)}$, if the interval $(\tau_1^{(q)}, \tau_1^{(q)} + \beta)$ contains strictly less than $k-1$ points (excluding the point $\tau_1^{(q)}$), we delete all the points in $(\tau_1^{(q)}, \tau_1^{(q)} + \beta)$. Otherwise, we leave the interval $(\tau_1^{(q)}, \tau_1^{(q)} + \beta)$ unchanged. Next, we repeat exactly the same procedure starting from the first point, say $\tau_i^{(q)}$, after the time $\tau_1^{(q)} + \beta$. Following the time $\tau_i^{(q)} + \beta$ we continue repeating this procedure indefinitely. The remaining (undeleted) points are enumerated in their increasing order as $\{\tau_n^{(d)}\}_{n\geq 1}$. Now, let T_U be defined as in (8) for the sequence $\{\tau_n^{(d)}\}_{n\geq 1}$ instead of $\{\tau_n^{(q)}\}_{n\geq 1}$. Since the sequence of points $\{\tau_n^{(d)}\}_{n\geq 1}$ is a subset of $\{\tau_n^{(q)}\}_{n\geq 1}$, it is clear that

$$T \leq T_U. \tag{10}$$

Now, we compute the distribution of T_U. Let X be a random variable independent of $\{\tau_n^{(q)}\}_{n\geq 1}$ with a geometric distribution $\mathbb{P}[X = i] = (1-p)^{i-1}p$, $i \geq 1$, where p is defined as

$$p \triangleq \mathbb{P}[M^{(q)}(0,\beta) \geq k-1].$$

Then, we claim that

$$T_U \stackrel{d}{=} \tau_X^{(q)} + \beta(X-1), \tag{11}$$

where $\stackrel{d}{=}$ represents equality in distribution. This equality follows from the construction of the sequence $\{\tau_n^{(d)}\}_{n\geq 1}$ and the memoryless property of the Poisson process. In this regard, if the interval $(\tau_1^{(q)}, \tau_1^{(q)} + \beta)$ contains more or equal to $(k-1)$ points, then $T_U = \tau_1^{(q)}$ and the probability of this event is $\mathbb{P}[M^{(q)}(\tau_1^{(q)}, \tau_1^{(q)} + \beta) \geq k-1] = \mathbb{P}[M^{(q)}(0,\beta) \geq k-1] = p$. Next, if $(\tau_1^{(q)}, \tau_1^{(q)} + \beta)$ contains less than $(k-1)$ points, then $T_U > \tau_1^{(q)} + \beta$ since we deleted all the points in $(\tau_1^{(q)}, \tau_1^{(q)} + \beta)$, and that happens with pobability $1-p$. Then, due to the memoryless property of the Poisson process, the first point $\tau_2^{(d)}$ of $\{\tau_i^{(d)}\}$ after $(\tau_1^{(q)} + \beta)$ is at an exponential distance from $(\tau_1^{(q)} + \beta)$. Furthermore, since the number of points of the Poisson process in non-intersecting intervals of the same length is independent and equally distributed, the probability that the interval $(\tau_2^{(d)}, \tau_2^{(d)} + \beta)$ contains more or equal to $(k-1)$ points of the process $\{\tau_n^{(q)}\}_{n\geq 1}$ is again p, and on this event $T_U = \tau_2^{(d)} \stackrel{d}{=} \tau_2^{(q)} + \beta$. Clearly, by repeating this argument one derives (11).

Next, since $\tau_X^{(q)}$ is the sum of X exponential random variables and X is independent of $\{\tau_n^{(q)}\}_{n\geq 1}$, then it is easy to show (see Theorem 5.3, p. 89 of [3]) that $\tau_X^{(q)}$ is also exponential with parameter pq. It is also straightforward to derive for any $\epsilon > 0$ and all $q \leq q_0 \equiv -\log(1-\epsilon/2)/\beta$

$$p = \mathbb{P}[M^{(q)}(0,\beta) \geq k-1] \geq \mathbb{P}[M^{(q)}(0,\beta) = k-1] = e^{-q\beta}\frac{(q\beta)^{k-1}}{(k-1)!} \geq (1-\epsilon/2)\frac{(q\beta)^{k-1}}{(k-1)!}. \tag{12}$$

At this point, using the observations from the previous paragraph, (10) and (11), we obtain, for all q small enough ($q \leq q_0$),

$$\begin{aligned} \mathbb{P}[T > t] &\leq \mathbb{P}[T_U > t] \\ &\leq \mathbb{P}\left[\tau_X^{(q)} > \left(1-\frac{\epsilon}{2}\right)t\right] + \mathbb{P}\left[\beta X > \frac{\epsilon}{2}t\right] &(13)\\ &\leq e^{-pq(1-\frac{\epsilon}{2})t} + (1-p)^{\frac{\frac{\epsilon}{2}t}{\beta}-1} &(14)\\ &\leq e^{-\frac{q^k\beta^{k-1}}{(k-1)!}(1-\epsilon)t} + 2e^{-h\epsilon q^{k-1}t}, \end{aligned}$$

where in the last inequality we applied the bound $1-x \leq e^{-x}, x \geq 0$ and assumed that q_0 is small enough such that $1/(1-p) \leq 2$. This completes the proof. $\diamond$

Next, we will prove the lower bound for the stopping time T defined in (8).

Lemma 3 *For any $\epsilon > 0$, there exists $q_0 > 0$ such that for all $0 < q \leq q_0$*

$$\mathbb{P}[T > t] \geq e^{-\frac{q^k\beta^{k-1}}{(k-1)!}(1+\epsilon)t}. \tag{15}$$

Proof: Since the bound is immediate for $k = 1$, we assume $k \geq 2$.

First, we relabel the points $\{\tau_n^{(q)}\}_{n\geq 1}$ as $\{\tau_n^{(q)}(i) \equiv \tau_n^{(q)}(i,\omega)\}$ using the following procedure. Let $\tau_1^{(q)}(0) = \tau_1^{(q)}$ and define a stopping time

$$Z_1 = \inf\{i \geq 1 : M^{(q)}(\tau_1^{(q)}(0) + (i-1)\beta, \tau_1^{(q)}(0) + i\beta) = 0\}.$$

Then, all the points in the interval $(\tau_1^{(q)}(0), \tau_1^{(q)}(0)+\beta Z_1)$ are labeled as $\tau_1^{(q)}(i), 1 \leq i \leq M^{(q)}(\tau_1^{(q)}(0), \tau_1^{(q)}(0)+\beta Z_1)$. Next, the first Poisson point after time $\tau_1^{(q)}(0) + \beta Z_1$ is named $\tau_2^{(q)}(0)$ and, similarly as before, we define a stopping time

$$Z_2 = \inf\{i \geq 1 : M^{(q)}(\tau_2^{(q)}(0) + (i-1)\beta, \tau_2^{(q)}(0) + i\beta) = 0\}.$$

Again, all the points in the interval $(\tau_2^{(q)}(0), \tau_2^{(q)}(0)+\beta Z_2)$ are labeled as $\tau_2^{(q)}(i), 1 \leq i \leq M^{(q)}(\tau_2^{(q)}(0), \tau^{(q)}(0)+\beta Z_2)$. We continue this procedure indefinitely. Note that, due to the Poisson memoryless property, the sequence of stopping times $\{Z_i\}$ is i.i.d. with geometric distribution

$$\mathbb{P}[Z_i = j] = (\mathbb{P}[M^{(q)}(0,\beta) > 0])^{j-1}\mathbb{P}[M^{(q)}(0,\beta) = 0]. \tag{16}$$

Next, we define for $n \geq 1$ sets

$$\mathcal{A}_n \triangleq \{\omega : \tau_n^{(q)}(i+k-1) - \tau_n^{(q)}(i) \leq \beta, 1 \leq j \leq M^{(q)}(\tau_n^{(q)}(0), \tau_n^{(q)} + \beta Z_i)\}.$$

Then, using the definition of $\tau_n^{(q)}(i)$, we show that

$$T = \inf\{\tau_n^{(q)}(i,\omega) : \omega \in \mathcal{A}_n, n \geq 1\} \geq T_L \triangleq \inf\{\tau_n^{(q)}(0,\omega) : \omega \in \mathcal{A}_n, n \geq 1\}, \tag{17}$$

where the equality follows from $|\tau_n^{(q)}(i) - \tau_m^{(q)}(j)| > \beta$, for any $n \neq m$, and the inequality is implied by $\tau_n^{(q)}(i) \geq \tau_n^{(q)}(0)$ for any $n \geq 1$.

Furthermore, we claim that

$$T_L \overset{d}{\geq} \nu_{X_L}, \tag{18}$$

where X_L is independent of $\{\tau_n^{(q)}\}$ and has geometric distribution $\mathbb{P}[X_L = j] = (1-p)^{j-1}p, j \geq 1$ with success probability

$$p = \mathbb{P}[\mathcal{A}_n] \leq \mathbb{P}[\{M^{(q)}(0,\beta) \geq k-1\} \cup \{M^{(q)}(0,\beta Z_1) \geq k\}]$$

(note that this p is different from the one in the proof of Lemma 2). The inequality in (18) follows from the memoryless property of the Possion process, the definition of Z_i and the observation that we can always reduce the value of the stopping time T_L by excluding the intervals of length βZ_i from its calculation. Furthermore, similarly as in the proof of Lemma 2, $\tau_{X_L}^{(q)}$ is an exponential random variable with distribution

$$\mathbb{P}[\tau_{X_L}^{(q)} > t] = e^{-pqt}. \tag{19}$$

Thus, in order to complete the proof, we need an upper bound on p. In this respect, using the union bound, we upper bound the success probability p as

$$\begin{aligned} p &\leq \mathbb{P}[\{M^{(q)}(0,\beta) \geq k-1\} \cup \{M^{(q)}(0,\beta Z_1) \geq k\}] \\ &\leq \mathbb{P}[M^{(q)}(0,\beta) \geq k-1] + \mathbb{P}[M^{(q)}(0,\beta Z_1) \geq k] \\ &\leq \mathbb{P}[M^{(q)}(0,\beta) \geq k-1] + \mathbb{P}[Z_1 > k] + \mathbb{P}[M^{(q)}(0,\beta k) \geq k] \\ &= \mathbb{P}[M^{(q)}(0,\beta) \geq k-1] + \mathbb{P}[M^{(q)}(0,\beta) > 0]^k + \mathbb{P}[M^{(q)}(0,\beta k) \geq k], \end{aligned} \tag{20}$$

where in the last equality we used the geometric distribution of Z_1 from (16). Finally, (17),(18), (19) and (20), in conjunction with

$$\mathbb{P}[M^{(q)}(0,\beta) \geq m] \leq \frac{(q\beta)^m}{m!} \sum_{i=0}^{\infty} (q\beta)^i \leq (1+\epsilon)\frac{(q\beta)^m}{m!}$$

for any $\epsilon > 0$ and all $q \leq \epsilon/\beta(1+\epsilon)$, yield the stated bound in the lemma. $\diamond$

3 Main results

In this section we derive our main results in Theorems 1, 2 and 3, where we estimate the asymptotics of the tail of the stationary search cost $C^{(N)}$ for α being greater, equal and less than one respectively. Our method of proof uses probabilistic and sample path arguments introduced in [11] for the case of ordinary LRU (PAC(β, 1)) algorithm. The starting point of our analysis is given by (7) derived in Section 2.

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.

In the following theorem, we assume that $N = \infty$ and denote $C \equiv C^{(N)}$; in the case of $N < \infty$, the proof would follow the analogous steps as in the proof of Theorem 2.

Theorem 1 *Assume that $q_i \sim c/i^\alpha$ as $i \to \infty$ and $\alpha > 1$. Then, as $x \to \infty$*

$$\mathbb{P}[C > x] \sim K_k(\alpha)\mathbb{P}[R > x], \tag{21}$$

where

$$K_k(\alpha) \triangleq \left[\Gamma\left(1 - \frac{1}{k\alpha}\right)\right]^{\alpha-1} \Gamma\left(1 + \frac{1}{k} - \frac{1}{k\alpha}\right). \tag{22}$$

Furthermore, function $K_k(\alpha)$ is monotonically increasing in α, for fixed k, with

$$\lim_{\alpha \downarrow 1} K_k(\alpha) = 1, \quad \lim_{\alpha \uparrow \infty} K_k(\alpha) = K_k(\infty) \triangleq \frac{1}{k}\Gamma\left(\frac{1}{k}\right) e^{\gamma/k}, \tag{23}$$

where γ is the Euler constant, i.e. $\gamma \approx 0.57721\ldots$, and monotonically decreasing in k, for fixed α, with

$$\lim_{k \to \infty} K_k(\alpha) = 1. \tag{24}$$

Remarks: (i) On Figure 1 we present drastic improvement in performance of the PAC(β, k), $k = 1, 2, 3,$ algorithm when compared to the optimal LFU policy. Note that even PAC($\beta, 3$) can not perform worse than 8% from the LFU algorithm, which implies near optimality of the PAC algorithm even for relatively small values of k; (ii) For the ordinary LRU policy (PAC(β, 1)), the convergence of $C^{(N)}$ in distribution as $N \to \infty$ and the Laplace transform of the limiting search cost C are first derived in [4]. Furthermore, the asymptotic result in the present form, using alternative approach that exploits the Tauberian technique for inverting the Laplace transform was originally derived in [8].

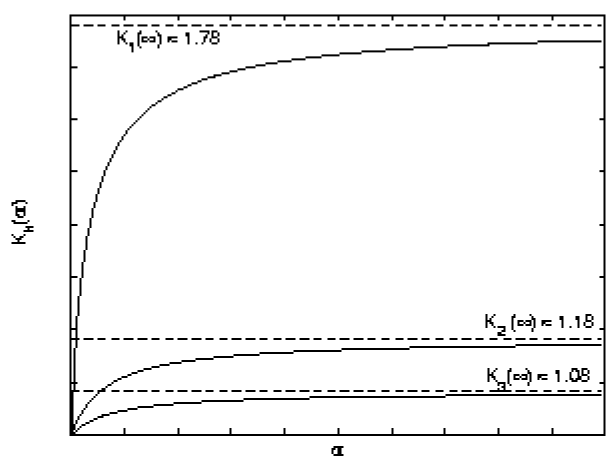

Figure 1: Function $K_k(\alpha)$

In the proofs of the following theorems we use results proved in Lemmas 4 and 5 of the Appendix.
Proof: First, we prove the upper bound for the asymptotic relationship in (21). Define the sum of indicator functions $S(t) \triangleq \sum_{j=1}^{\infty} 1[T_j < t]$; note that $S(t)$ is a.s. non-decreasing in t, i.e. $S(t) \leq S(t_0(x))$ a.s. for all $t \leq t_0(x)$. Then, after conditioning on T_i being larger or smaller than $t_0(x)$, the expression in (7) can be upper bounded as

$$\mathbb{P}[C > x] \leq \mathbb{P}[S(t_0(x)) > x] + \sum_{i=1}^{\infty} q_i \mathbb{P}[T_i \geq t_0(x)], \tag{25}$$

where in the previous expression we applied $\sum_{i=1}^{\infty} q_i = 1$ and $\mathbb{P}[S(t) > x] \leq 1$. Next, from the assumption of the theorem and Lemma 3 it follows that for any $\epsilon > 0$ there exists j_0 such that for all $j \geq j_0$ the bound (15) holds and therefore we can upper-bound the expectation of the sum $S(t)$ as

$$\begin{aligned} \mathbb{E}S(t) &= \sum_{j=1}^{\infty} \mathbb{P}[T_j < t] \\ &\leq j_0 + \sum_{j=1}^{\infty} \left(1 - e^{-\frac{q_j^k \beta^{k-1}}{(k-1)!}(1+\epsilon)t} \right). \end{aligned}$$

Next, using the preceding bound and Lemma 4 of the Appendix, we conclude that, as $t \to \infty$

$$\mathbb{E}S(t) \lesssim \Gamma\left(1 - \frac{1}{\alpha k}\right) \frac{c^{\frac{1}{\alpha}} \beta^{\frac{k-1}{\alpha k}}}{((k-1)!)^{\frac{1}{\alpha k}}} (1+\epsilon)^{\frac{1}{\alpha k}} t^{\frac{1}{\alpha k}}. \tag{26}$$

Now, if we select

$$t_0(x) = \frac{x^{\alpha k}(1-2\epsilon)^{\alpha k}(k-1)!}{(1+\epsilon)c^k \beta^{k-1}[\Gamma\left(1-\frac{1}{\alpha k}\right)]^{\alpha k}}$$

and use (26), it is easy to show that $\mathbb{E}S(t_0(x)) \leq (1-\epsilon)x$ for all x large enough. Now, large deviation bound for the sum of independent Bernoulli random variables proved in Lemma 5 of the Appendix implies

$$\mathbb{P}[S(t_0(x)) > x] \leq 2e^{-\theta \mathbb{E}S(t_0(x))},$$

for some $\theta > 0$. Thus, in conjunction with (25), we conclude that as $x \to \infty$

$$\mathbb{P}[C > x] \leq o\left(\frac{1}{x^{\alpha-1}}\right) + \sum_{i=1}^{\infty} q_i \mathbb{P}[T_i \geq t_0(x)]. \tag{27}$$

Next, from Lemma 2 there exists i_0 such that for all $i \geq i_0$, T_i satisfies (9). We use i_0 to denote a sufficiently large integer constant that is possibly different at different places in the proof. Then, since for every $i \leq i_0$, the inequality $q_i \geq q_{i_0}$ holds, the Poisson process of rate q_i can be constructed as a superposition of two independent Poisson processes with rates q_{i_0} and $q_i - q_{i_0}$. Therefore, in this construction, the process of rate q_i will have on each sample path more arrival points than the process of rate q_{i_0}. Thus, it is straightforward that the stopping time $T_i \leq T_{i_0}$ (a.s.), implying that for all $i \leq i_0$

$$\mathbb{P}[T_i \geq t] \leq \mathbb{P}[T_{i_0} \geq t]. \tag{28}$$

Therefore, we obtain

$$\begin{aligned} \sum_{i=1}^{\infty} q_i \mathbb{P}[T_i \geq t_0(x)] &\leq \sum_{i=1}^{i_0} q_i \mathbb{P}[T_{i_0} \geq t_0(x)] + \sum_{i=i_0}^{\infty} q_i e^{-\frac{q_i^k \beta^{k-1}(1-\epsilon)}{(k-1)!} t_0(x)} + \sum_{i=i_0}^{\infty} q_i e^{-h\epsilon q_i^{k-1} t_0(x)} \\ &\triangleq I_1(x) + I_2(x) + I_3(x), \end{aligned} \tag{29}$$

where in the last two sums we used the result of Lemma 2.

After using the bound (9) and replacing $t_0(x)$, it immediately follows that

$$I_1(x) \le \sum_{i=1}^{i_0} q_i \left[e^{-\frac{q_{i_0}^k \beta^{k-1}(1-\epsilon)}{(k-1)!} t_0(x)} + e^{-h\epsilon q_{i_0}^{k-1} t_0(x)} \right] = o\left(\frac{1}{x^{\alpha-1}}\right) \quad \text{as} \quad x \to \infty. \tag{30}$$

Now, by assumption of the theorem, for all i large enough ($i \ge i_0$, where i_0 is possibly larger than in (29))

$$(1-\epsilon)c/i^{\alpha} < q_i < (1+\epsilon)c/i^{\alpha}. \tag{31}$$

Furthermore, for i large enough ($i \ge i_0$) inequality $c/i^{\alpha} \le (1+\epsilon)c/u^{\alpha}$ holds for any $u \in [i, i+1]$ and, therefore, using this bound, (31), the monotonicity of the exponential function and replacing $t_0(x)$, yields

$$\begin{aligned} I_2(x) &\le (1+\epsilon) \sum_{i=i_0}^{\infty} \frac{c}{i^{\alpha}} e^{-\iota(\epsilon) \frac{x^{\alpha k}}{[\Gamma(1-\frac{1}{\alpha k})]^{\alpha k} i^{\alpha k}}} \\ &\le (1+\epsilon)^2 \int_1^{\infty} \frac{c}{u^{\alpha}} e^{-\iota(\epsilon) \frac{x^{\alpha k}}{[\Gamma(1-\frac{1}{\alpha k})]^{\alpha k} u^{\alpha k}}} du, \end{aligned} \tag{32}$$

where $\iota(\epsilon) \triangleq (1+\epsilon)^{-1}(1-\epsilon)^{k+1}(1-2\epsilon)^{\alpha k}$. Next, using the change of variable method for evaluating the integral with $z = x^{\alpha k}\iota(\epsilon)[\Gamma(1-\frac{1}{\alpha k})]^{-\alpha k} u^{-\alpha k}$, we obtain that the integral in (32) is equal to

$$\frac{c}{x^{\alpha-1}(\alpha-1)} \left[\Gamma\left(1-\frac{1}{\alpha k}\right)\right]^{\alpha-1} (\iota(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}} \frac{\alpha-1}{\alpha k} \int_0^{\frac{x^{\alpha k}\iota(\epsilon)}{[\Gamma(1-\frac{1}{\alpha k})]^{\alpha k}}} e^{-z} z^{\frac{1}{k}-\frac{1}{\alpha k}-1} dz,$$

which in conjunction with (32) yields

$$\limsup_{x\to\infty} \frac{I_2(x)}{\mathbb{P}[R > x]} \le K_k(\alpha)(\iota(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}} \to K_k(\alpha) \quad \text{as} \quad \epsilon \to 0, \tag{33}$$

where $K_k(\alpha)$ is defined in (22).

In order to estimate asymptotics of $I_3(x)$, we use analogous steps to those we applied in estimation of $I_2(x)$. Thus, from the assumption $q_i \sim c/i^{\alpha}$ as $i \to \infty$, it follows that for i large ($i \ge i_0$) inequalities (31) and $c/i^{\alpha} \le (1+\epsilon)c/u^{\alpha}$ hold for all $u \in [i, i+1]$ and, therefore, after replacing $t_0(x)$

$$\begin{aligned} I_3(x) &\le (1+\epsilon) \sum_{i=i_0}^{\infty} \frac{c}{i^{\alpha}} e^{-h\epsilon \frac{x^{\alpha k}}{i^{\alpha(k-1)}}} \\ &\le (1+\epsilon)^2 \int_1^{\infty} \frac{c}{u^{\alpha}} e^{-h\epsilon \frac{x^{\alpha k}}{u^{\alpha(k-1)}}} du. \end{aligned} \tag{34}$$

Now, if $k=1$, it is straightforward to compute the integral in the preceeding expression and obtain $I_3(x) \le (1+\epsilon)^2(c/(\alpha-1))e^{-h\epsilon x^{\alpha}} = o(1/x^{\alpha-1})$ as $x \to \infty$. Otherwise, for $k \ge 2$, after using the change of variable method for solving the integral in (34) where $z = h\epsilon x^{\alpha k} u^{-\alpha(k-1)}$, we obtain, as $x \to \infty$,

$$I_3(x) \le (1+\epsilon)^3 \frac{c}{(h\epsilon)^{\frac{1}{k-1}(1-\frac{1}{\alpha})}} \frac{1}{\alpha(k-1)} \frac{1}{x^{\frac{k}{k-1}(\alpha-1)}} \Gamma\left(\frac{1}{k-1} - \frac{1}{\alpha(k-1)}\right) = o\left(\frac{1}{x^{\alpha-1}}\right).$$

The previous expression, in conjunction with (33),(30), (29) and (27), yields, as $x \to \infty$

$$\mathbb{P}[C > x] \lesssim K_k(\alpha)\mathbb{P}[R > x]. \tag{35}$$

Next, we estimate an asymptotic lower bound for $\mathbb{P}[C > x]$ in (21). Similarly as before, Lemma 2 implies that for every $\epsilon > 0$ and all i large enough ($i \geq i_0$), inequality (9) holds and therefore

$$\begin{aligned}
\mathbb{E}S(t) &\geq \sum_{i=i_0}^{\lfloor Ht^{\frac{1}{\alpha k}} \rfloor} \left(1 - e^{-\frac{q_i^k \beta^{k-1}(1-\epsilon)}{(k-1)!}t} - e^{-h\epsilon q_i^{k-1} t} \right) \\
&\geq \sum_{i=i_0}^{\lfloor Ht^{\frac{1}{\alpha k}} \rfloor} \left(1 - e^{-\frac{q_i^k \beta^{k-1}(1-\epsilon)}{(k-1)!}t} \right) - \sum_{i=i_0}^{\lfloor Ht^{\frac{1}{\alpha k}} \rfloor} e^{-h\epsilon q_i^{k-1} t}. \qquad (36)
\end{aligned}$$

Next, since for i large ($i \geq i_0$) inequality (31) holds, after lower bounding the second sum in (36), we obtain, as $t \to \infty$

$$\begin{aligned}
\mathbb{E}S(t) &\geq \sum_{i=i_0}^{\lfloor Ht^{\frac{1}{\alpha k}} \rfloor} \left(1 - e^{-\frac{(1-\epsilon)^k c^k \beta^{k-1} (1-\epsilon) t}{i^{\alpha k}(k-1)!}} \right) - Ht^{\frac{1}{\alpha k}} e^{-h\epsilon \frac{(1-\epsilon)^{k-1} c^{k-1}}{H^{\alpha(k-1)}} t^{1-\frac{k-1}{k}}} \\
&= \sum_{i=i_0}^{\lfloor Ht^{\frac{1}{\alpha k}} \rfloor} \left(1 - e^{-\frac{(1-\epsilon)^{k+1} c^k \beta^{k-1} t}{i^{\alpha k}(k-1)!}} \right) + o(t^{\frac{1}{\alpha k}}) \\
&= \sum_{i=i_0}^{\infty} \left(1 - e^{-\frac{(1-\epsilon)^{k+1} c^k \beta^{k-1} t}{i^{k\alpha}(k-1)!}} \right) - \sum_{i=\lfloor Ht^{\frac{1}{\alpha k}} \rfloor + 1}^{\infty} \left(1 - e^{-\frac{(1-\epsilon)^{k+1} c^k \beta^{k-1} t}{i^{\alpha k}(k-1)!}} \right) + o(t^{\frac{1}{\alpha k}}). \qquad (37)
\end{aligned}$$

Now, after defining $L \triangleq c^k \beta^{k-1}(1-\epsilon)^{k+1}/(k-1)!$ and using the inequality $1 - e^{-x} \leq x$ for $x \geq 0$ small enough, we derive

$$\begin{aligned}
\lim_{t\to\infty} \frac{1}{t^{\frac{1}{\alpha k}}} \sum_{i=\lfloor Ht^{\frac{1}{\alpha k}} \rfloor + 1}^{\infty} \left(1 - e^{-\frac{c^k \beta^{k-1}(1-\epsilon)^{k+1} t}{i^{\alpha k}(k-1)!}} \right) &\leq \lim_{t\to\infty} \frac{1}{t^{\frac{1}{\alpha k}}} \int_{Ht^{\frac{1}{\alpha k}}}^{\infty} \frac{1}{u^{\alpha k}} Lt\,du \\
&= \frac{L}{\alpha k - 1} \frac{1}{H^{\alpha k - 1}} \to 0 \text{ as } H \to \infty, \qquad (38)
\end{aligned}$$

and, therefore, in conjunction with (37) and Lemma 4 of the Appendix, we conclude

$$\lim_{t\to\infty} \frac{\mathbb{E}S(t)}{t^{\frac{1}{\alpha k}}} \geq \Gamma\left(1 - \frac{1}{\alpha k}\right) \frac{c^{\frac{1}{\alpha}} \beta^{\frac{k-1}{\alpha k}}}{((k-1)!)^{\frac{1}{\alpha k}}} (1-\epsilon)^{\frac{k+1}{\alpha k}},$$

for any $\epsilon > 0$. Therefore, after letting $\epsilon \to 0$, we derive

$$\mathbb{E}S(t) \gtrsim \Gamma\left(1 - \frac{1}{\alpha k}\right) \frac{c^{\frac{1}{\alpha}} \beta^{\frac{k-1}{\alpha k}}}{((k-1)!)^{\frac{1}{\alpha k}}} t^{\frac{1}{\alpha k}} \text{ as } t \to \infty. \qquad (39)$$

Next, we redefine $t_0(x)$ as

$$t_0(x) \triangleq \frac{x^{\alpha k}(1+2\epsilon)^{\alpha k}(k-1)!}{\left[\Gamma(1-\frac{1}{\alpha k})\right]^{\alpha k} c^k \beta^{k-1}}.$$

Then, the monotonicity of $S(t)$ and (39) yield $\mathbb{E}S(t) \geq (1+\epsilon)x$ for all $t \geq t_0(x)$ and x large enough, which in conjunction with Lemma 5 of the Appendix implies $\mathbb{P}[S(t) > x] \geq 1 - \epsilon$ for all $t \geq t_0(x)$. Thus, for x

large enough, after conditioning on $T_i \geq t_0(x)$, expression (7) can be lower bounded as

$$\mathbb{P}[C > x] \geq (1-\epsilon) \sum_{i=1}^{\infty} q_i \mathbb{P}[T_i \geq t_0(x)]. \tag{40}$$

Next, we proceed with estimating (40). By inequality (15) of Lemma 3 and assumption of the theorem, for i large enough ($i \geq i_0$), after replacing $t_0(x)$, we obtain

$$\mathbb{P}[C > x] \geq (1-\epsilon) \sum_{i=i_0+1}^{\infty} q_i e^{-\frac{q_i^k x^{\alpha k}(1+\epsilon)(1+2\epsilon)^{\alpha k}}{[\Gamma(1-\frac{1}{\alpha k})]^{\alpha k} c^k}}.$$

Furthermore, similarly as in estimating $I_2(x)$, since for i large ($i \geq i_0$) inequalities (31) and $c/i^{\alpha} \geq (1-\epsilon)c/u^{\alpha}$ hold for all $u \in [i-1, i]$, which in conjunction with the monotonicity of the exponential function yields

$$\mathbb{P}[C > x] \geq (1-\epsilon)^3 \int_{u=i_0}^{\infty} \frac{c}{u^{\alpha}} e^{-\frac{x^{\alpha k}}{u^{\alpha k}[\Gamma(1-\frac{1}{\alpha k})]^{\alpha k}}\iota(\epsilon)} du, \tag{41}$$

where $\iota(\epsilon) \triangleq (1+\epsilon)^{k+1}(1+2\epsilon)^{\alpha k}$. Then, using the change of variable method for solving the previous integral with $z = x^{\alpha k}\iota(\epsilon)u^{-\alpha k}[\Gamma(1-\frac{1}{\alpha k})]^{-\alpha k}$, we obtain that the integral in (41) is equal to

$$\frac{c}{(\alpha-1)x^{\alpha-1}} \frac{\alpha-1}{\alpha k} (\iota(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}} \int_0^{\frac{x^{\alpha k}\iota(\epsilon)}{i_0^{\alpha k}[\Gamma(1-\frac{1}{\alpha k})]^{\alpha k}}} z^{\frac{1}{k}-\frac{1}{\alpha k}-1} e^{-z} dz,$$

which implies

$$\liminf_{x\to\infty} \frac{\mathbb{P}[C > x]}{\mathbb{P}[R > x]} \geq (\iota(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}} K_k(\alpha) \to K_k(\alpha) \;\text{ as }\; \epsilon \to 0,$$

where $K_k(\alpha)$ is defined in (22). The previous result in conjunction with (35) proves (21).

Finally, it is left to prove the monotonicity of function $K_k(\alpha)$ and its limits when $\alpha \to \infty$, $k \to \infty$. Since the function on the right hand side of (22) is well-defined for all real $k \geq 1$, we will assume that $K_k(\alpha)$ is defined for all real values of $k \geq 1$ and $\alpha > 1$ as well. Thus, proving that $K_k(\alpha)$ is monotonic over real k will in particular imply the monotonicity over integer values.

First, we prove the monotonicity in α for all fixed k. Since PAC(β,1) algorithm corresponds to the case of the ordinary LRU, their performances are the same and the monotonicity of $K_1(\alpha)$ follows from Theorem 3 of [8]. Thus, we continue with the proof of monotonicity for values of $k \geq 2$. Proving that $K_k(\alpha)$ is monotonically increasing in α is equivalent to showing that $\log K_k(\alpha)$ monotonically increases in α. Thus, we observe the following function

$$\begin{aligned} \log K_k(\alpha) &= \log \frac{\alpha-1}{\alpha k} + (\alpha-1)\log\Gamma\left(1-\frac{1}{k\alpha}\right) + \log\Gamma\left(\frac{1}{k}-\frac{1}{k\alpha}\right) \\ &= \log\Gamma\left(1+\frac{1}{k}-\frac{1}{\alpha k}\right) + (\alpha-1)\log\Gamma\left(2-\frac{1}{\alpha k}\right) - (\alpha-1)\log\left(1-\frac{1}{\alpha k}\right). \end{aligned} \tag{42}$$

Function $\log K_k(\alpha)$ monotonically increases in α if its derivative with respect to α is positive for all $\alpha > 1$. Thus,

$$\begin{aligned} \frac{d}{d\alpha}(\log K_k(\alpha)) = &\log\Gamma\left(2-\frac{1}{k\alpha}\right) - \log\left(1-\frac{1}{k\alpha}\right) - \frac{\alpha-1}{\alpha}\frac{1}{k\alpha-1} \\ &+ \frac{\alpha-1}{\alpha^2 k}\Psi^{(0)}\left(2-\frac{1}{k\alpha}\right) + \frac{1}{\alpha^2 k}\Psi^{(0)}\left(1+\frac{1}{k}-\frac{1}{k\alpha}\right), \end{aligned} \tag{43}$$

where $\Psi^{(k)}$, $k = 0, 1, \ldots$, are Polygamma functions (see equation 6.4.1, p. 260 of [1]). Furthermore, since $\Psi^{(0)}(1) = -\gamma$ (Euler's constant) and $\Gamma(1) = 1$, by using continuity and passing $\alpha \to \infty$ in (43), we conclude

$$\frac{d}{d\alpha} \log K_k(\alpha) \to 0 \text{ as } \alpha \to \infty. \tag{44}$$

Therefore, to show that $d/d\alpha \log K_k(\alpha) \geq 0$ for all $\alpha > 1$, it is enough to prove that the second derivative

$$\begin{aligned} \frac{d^2}{d\alpha^2} \log K_k(\alpha) =& \frac{1}{k^2\alpha^4}\left[\Psi^{(1)}\left(1 + \frac{1}{k} - \frac{1}{k\alpha}\right) + (\alpha - 1)\Psi^{(1)}\left(2 - \frac{1}{k\alpha}\right)\right] \\ &+ \frac{2}{k\alpha^3}\left[\Psi^{(0)}\left(2 - \frac{1}{\alpha k}\right) - \Psi^{(0)}\left(1 + \frac{1}{k} - \frac{1}{k\alpha}\right)\right] \end{aligned} \tag{45}$$

is non-positive for all $\alpha \in (1, \infty)$. Now, equation 6.4.1 on page 260 of [1] implies that $\Psi^{(0)}(z)$ is monotonically increasing ($\Psi^{(1)}(z) \geq 0$) and concave ($\Psi^{(2)}(z) \leq 0$) and, similarly, $\Psi^{(1)}(z)$ is monotonically decreasing ($\Psi^{(2)}(z) \leq 0$) and convex ($\Psi^{(3)}(z) \geq 0$) for all $z > 0$. Thus, since for any $\alpha > 1$, $k \geq 1$, arguments of the functions $\Psi^{(0)}$ and $\Psi^{(1)}$, i.e. $2 - 1/(k\alpha)$ and $1 + 1/k - 1/(k\alpha)$, belong to the interval $[1, 2]$, and given the values $\Psi^{(0)}(1) = -\gamma$, $\Psi^{(0)}(2) = -\gamma + 1$, $\Psi^{(1)}(1) = \pi^2/6$, $\Psi^{(1)}(2) = \pi^2/6 - 1$, we derive the following linear bounds for all $1 \leq z \leq 2$:

$$\Psi^{(0)}(z) \geq -\gamma - 1 + z \quad \text{and} \quad \Psi^{(1)}(z) \leq \frac{\pi^2}{6} + 1 - z. \tag{46}$$

Next, by first upper bounding $\Psi^{(0)}(2 - 1/(k\alpha))$ with $1 - \gamma$, and then using the inequalities from (46) in (45), one obtains after some easy algebra

$$\frac{d^2}{d\alpha^2} \log K_k(\alpha) \leq \frac{18 + (-18 - 24k + \pi^2)\alpha - 2k(-18 + \pi^2)\alpha^2 + k^2(-12 + \pi^2)\alpha^3}{6k^2\alpha^4(\alpha k - 1)^2}. \tag{47}$$

Thus, it is left to prove that for any $\alpha > 1$ and $k \geq 2$, the expression on the right hand side of (47) is less or equal to zero. In that respect, we analyze the numerator of (47)

$$f(\alpha, k) \triangleq 18 + (-18 - 24k + \pi^2)\alpha - 2k(-18 + \pi^2)\alpha^2 + k^2(-12 + \pi^2)\alpha^3.$$

First, we show that function $f(\alpha, k)$ is decreasing in k for all $k \geq 2$ and any fixed $\alpha > 1$. Thus, by taking the derivative of f with respect to k, we obtain

$$\frac{df}{dk} = \alpha[2k\alpha^2(-12+\pi^2) - 2\alpha(-18+\pi^2) - 24] \leq \alpha[2\alpha^2(-24+2\pi^2) - 2\alpha(-18+\pi^2) - 24] < -8\alpha < -8,$$

since the maximum of the quadratic term in the previous expression is less than -8. Therefore, function $f(\alpha, k)$ decreases in k. Since $k \geq 2$ and $\alpha > 1$, we can upper-bound its value by $f(\alpha, 2)$, i.e.

$$\begin{aligned} f(\alpha, k) &\leq 18(1 - \alpha) + (-48 + \pi^2)\alpha - 4(-18 + \pi^2)\alpha^2 + 4(-12 + \pi^2)\alpha^3 \\ &\leq \alpha[4\alpha^2(-12 + \pi^2) + \alpha(72 - 4\pi^2) - 48 + \pi^2] < -7\alpha < -7, \end{aligned}$$

since in this case the quadratic term in the second inequality is less than -7. Therefore, we obtained that $f(\alpha, k) \leq 0$ for all $k \geq 2$, $\alpha > 1$, and, thus, from (47) it follows that $d^2/d\alpha^2 \log K_k(\alpha) \leq 0$. This concludes the proof of monotonicity of function $K_k(\alpha)$ in α. Finally, the second limit in (23) follows by straightforward application of the equation 6.1.33, p. 256 of [1]. Next, the first limit in (23) in the case of

$k = 1$ follows from Theorem 3 of [8]. Otherwise, for $k \geq 2$, the limit follows directly from expression (22) after replacing $\alpha = 1$.

In order to prove the monotonicity of function $K_k(\alpha)$ in k, we observe the first derivative with respect to k of function $\log K_k(\alpha)$ and obtain

$$\frac{d}{dk}\log K_k(\alpha) = \left(1 - \frac{1}{\alpha}\right)\frac{1}{k^2}\left[\Psi^{(0)}\left(1 - \frac{1}{k\alpha}\right) - \Psi^{(0)}\left(1 + \frac{1}{k} - \frac{1}{k\alpha}\right)\right] < 0, \tag{48}$$

where the last inequality follows from the monotonicity of function $\Psi^{(0)}(z)$ as discussed before. Thus, $K_k(\alpha)$ is monotonically decreasing in k for all real $k \geq 1$. In particular, $K_k(\alpha)$ is decreasing for integer values of $k \geq 1$. Finally, the monotonicity of $K_k(\alpha)$ and (23) imply (24), i.e.

$$1 = K_k(1) \leq K_k(\alpha) \leq K_k(\infty) \to 1 \text{ as } k \to \infty,$$

which concludes the proof of the theorem.

◇

Theorem 2 *Assume that $q_i = h_N/i^\alpha$, $1 \leq i \leq N$, where h_N is the normalization constant and $0 < \alpha < 1$. Then, for any $0 < \delta < 1$, as $N \to \infty$*

$$\mathbb{P}[C^{(N)} > \delta N] \sim F_k(\delta) \triangleq \frac{1-\alpha}{\alpha k}(\eta_\delta)^{\frac{1}{\alpha k} - \frac{1}{k}}\Gamma\left(\frac{1}{k} - \frac{1}{\alpha k}, \eta_\delta\right), \tag{49}$$

where η_δ is the unique solution of the equation

$$1 - \frac{1}{\alpha k}\Gamma\left(-\frac{1}{\alpha k}, \eta\right)\eta^{\frac{1}{\alpha k}} = \delta;$$

note that $\Gamma(x, y)$, $y > 0$, is the incomplete Gamma function, i.e. $\Gamma(x, y) = \int_y^\infty e^{-t}t^{x-1}dt$. Furthermore, $F_k(\delta)$, $\delta \in (0, 1)$, is a proper distribution, with $\lim_{\delta\to 0} F_k(\delta) = 1$, $\lim_{\delta\to 1} F_k(\delta) = 1$ and

$$\lim_{k\to\infty} F_k(\delta) = 1 - \delta^{1-\alpha}. \tag{50}$$

Remarks: (i) On Figure 2 we present the relative performance of the PAC$(20, k)$ replacement scheme, $k = 1, 2, 3$, when compared to the optimal static arrangement; in this case the cache can store $1/10$ of the total number of documents. Note that the performance of the PAC algorithm drastically improves even for the small values of parameter k, and, thus, we can claim its near optimality; (ii) For the ordinary MTF searching ($k = 1$), the convergence of $C^{(N)}/N$ in distribution as $N \to \infty$ and the Laplace transform of the limiting function are obtained in Lemma 4.5 of [4]. The result in its presented form, also for the ordinary LRU, was derived in [9]; (iii) It is possible to relax the assumption $q_i = h_N/i^\alpha$, $1 \leq i \leq N$, and that would involve some minor changes in the proof of this result. For example, by assuming that for any $\epsilon > 0$, there exists i_0, such that for all $i_0 \leq i \leq N$, inequality $(1-\epsilon)c/(i^\alpha N^{1-\alpha}) \leq q_i \leq (1+\epsilon)c/(i^\alpha N^{1-\alpha})$ holds, the expression in (53) would be replaced by this, slightly different form, and the rest of the proof would be identical; the final asymptotic formula in this case would have factor c instead of $1 - \alpha$ and all of the other factors in (49) would be the same.

Proof: First, we estimate the upper bound for the asymptotic relationship in (49). Similarly as in the proof of Theorem 1, we define the sum $S(t) \triangleq \sum_{j=1}^N 1[T_j < t]$. For any $t_0 > 0$, since $S(t)$ is (a.s.) non-decreasing

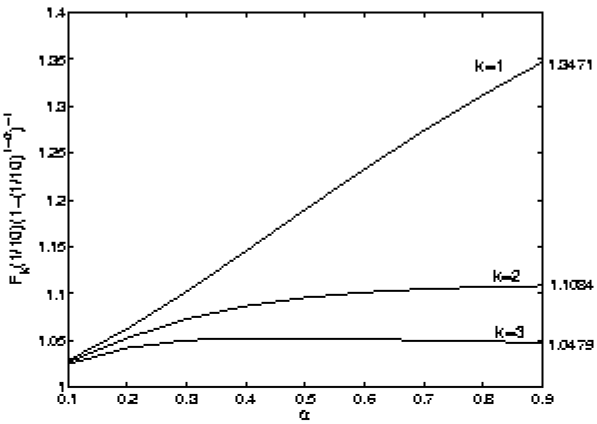

Figure 2: Ratio $F_k(1/10)/(1-(1/10)^{1-\alpha})$

in t, we have $S(t) \le S(t_0)$ a.s. for all $t \le t_0$. Thus, after conditioning on T_i being larger or smaller than t_0 in expression (7), we easily obtain the following upper bound

$$\mathbb{P}[C^{(N)} > \delta N] \le \mathbb{P}[S(t_0) > \delta N] + \sum_{i=1}^{N} q_i \mathbb{P}[T_i \ge t_0]. \tag{51}$$

Then, by Lemma 3, for N large and any $\epsilon > 0$, there exists i_0 such that for all $i \ge i_0$ inequality (15) holds, and, in conjunction with the monotonicity of the exponential function, we obtain

$$\begin{aligned}
\mathbb{E}S(t) &= \sum_{i=1}^{N} \mathbb{P}[T_i < t] \\
&\le i_0 + \sum_{i=i_0+1}^{N} \left(1 - e^{-q_i^k \frac{\beta^{k-1}(1+\epsilon)t}{(k-1)!}}\right) \\
&\le N - \int_{i_0}^{N} e^{-\frac{h_N^k}{u^{\alpha k}} \frac{\beta^{k-1}}{(k-1)!}(1+\epsilon)t} du.
\end{aligned} \tag{52}$$

At this point, note that from the assumption $\sum_{i=1}^{N} h_N/i^{\alpha} = 1$, it is straightforward to check that for N large enough

$$(1-\epsilon)\frac{1-\alpha}{N^{1-\alpha}} \le h_N \le (1+\epsilon)\frac{1-\alpha}{N^{1-\alpha}}. \tag{53}$$

Thus, if we assume that t increases in N as $t = \eta N^k/\xi(\epsilon)$, where $\eta > 0$ is a fixed constant, using inequalities (52) and (53), we bound $\mathbb{E}S(\eta N^k/\xi(\epsilon))$ for N large as

$$\mathbb{E}S\left(\frac{\eta N^k}{\xi(\epsilon)}\right) \le N - \int_{i_0}^{N} e^{-\frac{N^{\alpha k}\eta}{u^{\alpha k}}} du, \tag{54}$$

where $\xi(\epsilon)$ is defined as

$$\xi(\epsilon) \triangleq \frac{(1+\epsilon)^{k+1}(1-\alpha)^k \beta^{k-1}}{(k-1)!}.$$

Next, using the change of variable $z = N^{\alpha k}\eta u^{-\alpha k}$ method in the integral in (54), we obtain

$$\mathbb{E}S\left(\frac{\eta N^k}{\xi(\epsilon)}\right) \le N - N\frac{1}{\alpha k}\eta^{\frac{1}{\alpha k}} \int_{\eta}^{\eta\frac{N^{\alpha k}}{t_0^{\alpha k}}} e^{-z} z^{-1-\frac{1}{\alpha k}} dz. \tag{55}$$

Furthermore, since

$$\int_{\eta}^{\eta N^{k\alpha}} e^{-z} z^{-1-\frac{1}{\alpha k}} dz \uparrow \Gamma\left(-\frac{1}{\alpha k}, \eta\right) \text{ as } N \to \infty,$$

we obtain, for N large enough

$$\mathbb{E}S\left(\frac{\eta N^k}{\xi(\epsilon)}\right) \le N - N(1-\epsilon)\frac{1}{\alpha k}\eta^{\frac{1}{\alpha k}}\Gamma\left(-\frac{1}{\alpha k}, \eta\right). \tag{56}$$

Now, define the function $f(\eta)$ as

$$f(\eta) \triangleq 1 - \frac{1}{\alpha k}\eta^{\frac{1}{\alpha k}}\Gamma\left(-\frac{1}{\alpha k}, \eta\right), \tag{57}$$

and let $\eta_\delta(\epsilon_0)$ be a unique solution to the equation

$$f(\eta) = \delta(1 - 2\epsilon_0), \tag{58}$$

for some $\epsilon_0 > 0$; the uniqueness of the solution will be justified at the end of the proof. Then, from (56), it follows

$$\mathbb{E}S\left(\frac{\eta_\delta(\epsilon_0)N^k}{\xi(\epsilon)}\right) \le N(1 - (1-\epsilon)(1 - \delta(1-2\epsilon_0))).$$

Thus, for all $\epsilon < \epsilon_0\delta(1-2\epsilon_0)/(1-\delta(1-2\epsilon_0))$, the preceding expressions are bounded ny

$$\mathbb{E}S\left(\frac{\eta_\delta N^k}{\xi(\epsilon)}\right) \le (1+\epsilon_0)\delta(1-2\epsilon_0)N \le (1-\epsilon_0)\delta N.$$

Now, since $S(t)$ is a.s. non-decreasing in t, we conclude that for all $t \le t_0 \triangleq \eta_\delta(\epsilon_0)N^k(\xi(\epsilon))^{-1}$

$$\mathbb{E}S(t) \le (1-\epsilon_0)\delta N.$$

At this point, we use the large deviation (Chernoff) bound for the sum of N independent Bernoulli random variables from Lemma 5 of the Appendix

$$\mathbb{P}[S(t_0) > \delta N] \le e^{-\theta_{\epsilon_0}\delta N},$$

for some $\theta_{\epsilon_0} > 0$, and, thus, after upper bounding the first term in (51) and replacing t_0, we derive

$$\mathbb{P}[C^{(N)} > \delta N] \le o(1) + \sum_{i=1}^{N} q_i\mathbb{P}[T_i \ge t_0] \text{ as } N \to \infty. \tag{59}$$

Next, we estimate the second term on the right hand side of (59). The same arguments that resulted in inequality (28) in the proof of Theorem 1, in conjunction with (9), yield, for N and i_0 large enough,

$$\begin{aligned}
\sum_{i=1}^{N} q_i\mathbb{P}[T_i \ge t_0] &\le \sum_{i=1}^{i_0} q_i\mathbb{P}[T_i \ge t_0] + \sum_{i=i_0}^{N} q_i\mathbb{P}[T_i \ge t_0] \\
&\le i_0\mathbb{P}[T_{i_0} \ge t_0] + \sum_{i=i_0}^{N} q_i e^{-\frac{q_i^k\beta^{k-1}(1-\epsilon)}{(k-1)!}t_0} + \sum_{i=i_0}^{N} q_i e^{-h\epsilon q_i^{k-1}t_0} \\
&\triangleq I_1(\delta) + I_2(\delta) + I_3(\delta).
\end{aligned} \tag{60}$$

After replacing t_0 in $I_1(\delta)$, it is straightforward to conclude

$$I_1(\delta) \leq i_0 \left[e^{-\frac{q_{i_0}^k \beta^{k-1}(1-\epsilon)}{(k-1)!} t_0} + e^{-h\epsilon q_{i_0}^{k-1} t_0} \right] = o(1) \text{ as } N \to \infty. \tag{61}$$

Next, we estimate $I_2(\delta)$. After applying inequality (53), replacing t_0, using the monotonicity of the exponential function and relation $h_N/i^\alpha \leq (1+\epsilon)h_N/u^\alpha$ for all $u \in [i, i+1]$ and i large ($i \geq i_0$), we obtain

$$I_2(\delta) \leq (1+\epsilon)^2 \int_{i_0}^{N} \frac{1-\alpha}{u^\alpha N^{1-\alpha}} e^{-\iota(\epsilon)\frac{\eta_\delta(\epsilon_0)}{u^{\alpha k}} N^{\alpha k}} du, \tag{62}$$

where we define $\iota(\epsilon) \triangleq (1-\epsilon)^{k+1}(1+\epsilon)^{-(k+1)}$. Then, similarly as before, using the change of variable method for solving the integral with $z \triangleq \iota(\epsilon)\eta_\delta(\epsilon_0)u^{-\alpha k}N^{\alpha k}$, we derive

$$\begin{aligned} I_2(\delta) &\leq (1+\epsilon)^2 (\iota(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}} \frac{1-\alpha}{\alpha k} (\eta_\delta(\epsilon_0))^{\frac{1}{\alpha k}-\frac{1}{k}} \int_{\iota(\epsilon)\eta_\delta(\epsilon_0)}^{\iota(\epsilon)\frac{\eta_\delta(\epsilon_0)N^{\alpha k}}{i_0^{\alpha k}}} z^{\frac{1}{k}-\frac{1}{\alpha k}-1} e^{-z} dz \\ &\leq (1+\epsilon)^2 (\iota(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}} \frac{1-\alpha}{\alpha k} (\eta_\delta(\epsilon_0))^{\frac{1}{\alpha k}-\frac{1}{k}} \Gamma\left(\frac{1}{k} - \frac{1}{\alpha k}, \iota(\epsilon)\eta_\delta(\epsilon_0)\right). \end{aligned}$$

Thus, after letting $\epsilon \downarrow 0, \epsilon_0 \downarrow 0$, we conclude

$$\limsup_{N\to\infty} I_2(\delta) \leq \frac{1-\alpha}{\alpha k} (\eta_\delta)^{\frac{1}{\alpha k}-\frac{1}{k}} \Gamma\left(\frac{1}{k} - \frac{1}{\alpha k}, \eta_\delta\right), \tag{63}$$

since, by continuity of $f(\eta)$, $\eta_\delta(\epsilon_0) \to \eta_\delta$ as $\epsilon_0 \downarrow 0$, where η_δ is the unique solution to the equation $f(\eta_\delta) = \delta$.

Finally, we estimate $I_3(\delta)$. Here, we observe two possible cases: $k = 1$ and $k \geq 2$. In the case when $k = 1$, after applying (53), replacing t_0, and using relation $h_N/i^\alpha \leq (1+\epsilon)h_N/u^\alpha$ for all $u \in [i, i+1]$ and i large ($i \geq i_0$), we obtain

$$\begin{aligned} I_3(\delta) &\leq (1+\epsilon)^2 \frac{1-\alpha}{N^{1-\alpha}} e^{-h\epsilon N^k} \int_{i_0}^{N} \frac{1}{u^\alpha} du \\ &\leq (1+\epsilon)^2 e^{-h\epsilon N^k} = o(1) \text{ as } N \to \infty. \end{aligned}$$

Next, we estimate $I_3(\delta)$ in the case when $k \geq 2$. Similarly as before, after using (53), replacing t_0, in conjunction with the monotonicity of the exponential function and $h_N/i^\alpha \leq (1+\epsilon)h_N/u^\alpha$ for all $u \in [i, i+1]$ and i large ($i \geq i_0$), we obtain

$$I_3(\delta) \leq \frac{(1+\epsilon)^2(1-\alpha)}{N^{1-\alpha}} \int_{i_0}^{N} \frac{1}{u^\alpha} e^{-\frac{h\epsilon N^{\alpha k-\alpha+1}}{u^{\alpha(k-1)}}} du.$$

Then, using the change of variable method for solving the integral, with $z = h\epsilon N^{\alpha k-\alpha+1}u^{-\alpha(k-1)}$, we derive, for N large,

$$\begin{aligned} I_3(\delta) &\leq h\epsilon^{\frac{1}{k-1}(\frac{1}{\alpha}-1)} N^{\frac{1}{k-1}(\frac{1}{\alpha}-1)} \int_{h\epsilon N}^{h\epsilon N(\frac{N}{i_0})^{\alpha(k-1)}} e^{-z} z^{-\frac{1}{k-1}(\frac{1}{\alpha}-1)-1} dz \\ &\leq h\epsilon^{\frac{1}{k-1}(\frac{1}{\alpha}-1)} N^{\frac{1}{k-1}(\frac{1}{\alpha}-1)} \int_{h\epsilon N}^{\infty} e^{-z} dz \\ &\leq h\epsilon^{\frac{1}{k-1}(\frac{1}{\alpha}-1)} N^{\frac{1}{k-1}(\frac{1}{\alpha}-1)} e^{-h\epsilon N} = o(1) \text{ as } N \to \infty. \end{aligned} \tag{64}$$

Finally, (64), (63), (61), (60) and (59) imply

$$\limsup_{N\to\infty} \mathbb{P}[C^{(N)} > \delta N] \le \frac{1-\alpha}{\alpha k}(\eta_\delta)^{\frac{1}{\alpha k}-\frac{1}{k}}\Gamma\left(\frac{1}{k}-\frac{1}{\alpha k}, \eta_\delta\right). \tag{65}$$

Next, we estimate the asymptotic lower bound for $\mathbb{P}[C^{(N)} > \delta N]$ in (49). By Lemma 2, for any $\epsilon > 0$ and N large, there exists i_0 such that for all $i \ge i_0$ inequality (9) holds, and, in conjunction with the monotonicity of the exponential function, we obtain

$$\begin{aligned}
\mathbb{E}S(t) &= \sum_{i=1}^{N} \mathbb{P}[T_i < t] \ge \sum_{i=i_0}^{N} \mathbb{P}[T_i < t] \\
&\ge \sum_{i=i_0}^{N} \left(1 - e^{-\frac{q_i^k \beta^{k-1}(1-\epsilon)}{(k-1)!}t} - e^{-h\epsilon q_i^{k-1} t}\right) \\
&\ge N - i_0 - \int_{i_0}^{N} e^{-\frac{k_N^k}{u^{\alpha k}}\frac{\beta^{k-1}(1-\epsilon)}{(k-1)!}t} du - \int_{i_0}^{N} e^{-h\epsilon \frac{k_N^{k-1}}{u^{\alpha(k-1)}}t} du.
\end{aligned}$$

Then, after replacing the bound in (53), letting t increase in N as $t = \eta N^k/\xi(\epsilon)$, where, similarly as before, $\eta > 0$ is a fixed constant, we define

$$\begin{aligned}
\mathbb{E}S\left(\frac{\eta N^k}{\xi(\epsilon)}\right) &\ge N - i_0 - \int_{i_0}^{N} e^{-\frac{N^{\alpha k}\eta}{u^{\alpha k}}} du - \int_{i_0}^{N} e^{-h\epsilon \frac{N^{\alpha k-\alpha+1}}{u^{\alpha k-\alpha}}} du \\
&\triangleq N - i_0 - I_1 - I_2,
\end{aligned} \tag{66}$$

with $\xi(\epsilon)$ redefined as $\xi(\epsilon) \triangleq (1-\epsilon)^{k+1}\beta^{k-1}(1-\alpha)^k((k-1)!)^{-1}$.

First, we estimate the upper bound of I_1 for N large. Using completely analogous steps to those applied in estimating expressions (54), (55) and (56), after using the change of variable method to solve the integral with $z = N^{\alpha k}\eta u^{-\alpha k}$, we obtain

$$\begin{aligned}
I_1 &\le N\eta^{\frac{1}{\alpha k}} \frac{1}{\alpha k} \int_{\eta}^{(\frac{N}{i_0})^{\alpha k}\eta} e^{-z} z^{-1-\frac{1}{\alpha k}} dz \\
&\le N\eta^{\frac{1}{\alpha k}} \frac{1}{\alpha k} \Gamma\left(-\frac{1}{\alpha k}, \eta\right).
\end{aligned} \tag{67}$$

Now, we redefine $\eta_\delta(\epsilon)$ to be the unique solution to the equation $f(\eta_\delta) = (1+2\epsilon)\delta$, for some $\epsilon > 0$, where $f(\eta)$ was defined in (57). Then, after replacing $\eta_\delta(\epsilon)$ in (67), we obtain

$$I_1 \le N\eta_\delta(\epsilon)^{\frac{1}{\delta}} \frac{1}{\alpha k}\Gamma\left(-\frac{1}{\alpha k}, \eta_\delta(\epsilon)\right) = N(1 - \delta(1+2\epsilon)). \tag{68}$$

Next, we estimate the upper bound of I_2. Similarly as in estimating $I_3(\delta)$ before, we observe two possible cases: $k = 1$ and $k \ge 2$. In the case where $k = 1$, we obtain

$$I_2 \le \int_{i_0}^{N} e^{-h\epsilon N} du = (N - i_0)e^{-h\epsilon N} = o(1) \text{ as } N \to \infty.$$

Furthermore, in the case $k \geq 2$, after using the change of variable method for solving the integral, where $z = h\epsilon N^{\alpha k - \alpha + 1} u^{-\alpha k + \alpha}$, we compute, for N large

$$
\begin{aligned}
I_2 &\leq \int_{i_0}^{N} e^{-\frac{h\epsilon N^{\alpha k - \alpha + 1}}{u^{\alpha(k-1)}}} du \\
&\leq h\epsilon^{\frac{1}{\alpha(k-1)}} N^{1+\frac{1}{\alpha(k-1)}} \int_{h\epsilon N}^{\infty} e^{-z} dz \\
&\leq h\epsilon^{\frac{1}{\alpha(k-1)}} N^{1+\frac{1}{\alpha(k-1)}} e^{-h\epsilon N} = o(1) \ \text{ as } \ N \to \infty.
\end{aligned} \tag{69}
$$

Finally, (69), (68) and (66) imply that for any $\epsilon > 0$ and N large enough

$$
\mathbb{E}S\left(\frac{\eta_\delta(\epsilon) N^k}{\xi(\epsilon)}\right) \geq N - i_0 - N(1 - \delta(1 + 2\epsilon)) - \epsilon,
$$

which for all $N \geq (i_0 + \epsilon + \delta(1+\epsilon))/(1+2\epsilon)\delta$ yields

$$
\mathbb{E}S\left(\frac{\eta_\delta(\epsilon) N^k}{\xi(\epsilon)}\right) \geq (1+\epsilon)\delta. \tag{70}
$$

Now, since $S(t)$ is a.s. increasing in t, we obtain that for all N and t large, $t > t_0 \triangleq \eta_\delta(\epsilon) N^k (\xi(\epsilon))^{-1}$,

$$
\mathbb{E}S(t) \geq (1+\epsilon)\delta N. \tag{71}
$$

At this point, using the previous observations, the monotonicity of $S(t)$ and (7), after conditioning on T_i being greater than t_0, we obtain the lower bound

$$
\mathbb{P}[C^{(N)} > \delta N] \geq \mathbb{P}[S(t_0) > \delta N] \sum_{i=1}^{N} q_i \mathbb{P}[T_i \geq t_0]. \tag{72}
$$

Now, given the inequality (71), the large deviation (Chernoff) bound from Lemma 5 of the Appendix implies for any $\epsilon > 0$ and N large enough

$$
\mathbb{P}[S(t_0) > \delta N] \geq 1 - \epsilon.
$$

Thus, the previous inequality and (72) yield

$$
\mathbb{P}[C^{(N)} > \delta N] \geq (1-\epsilon) \sum_{i=1}^{N} q_i \mathbb{P}[T_i \geq t_0].
$$

Now, by Lemma 3, for any $\epsilon > 0$ and i large ($i \geq i_0$), inequality (15) holds, and, therefore

$$
\mathbb{P}[C^{(N)} > \delta N] \geq (1-\epsilon) \sum_{i=i_0+1}^{N} \frac{h_N}{i^\alpha} e^{-\frac{h_N^k \beta^{k-1}(1+\epsilon)t_0}{i^{\alpha k}(k-1)!}},
$$

which, using the inequality $h_N/i^\alpha \geq (1-\epsilon) h_N/u^\alpha$ for all i large ($i \geq i_0$) and $u \in [i-1, i]$, the monotonicity of the exponential function and inequality (53), after replacing t_0, implies

$$
\mathbb{P}[C^{(N)} > \delta N] \geq (1-\epsilon)^3 \frac{1-\alpha}{N^{1-\alpha}} \int_{i_0}^{N} \frac{1}{u^\alpha} e^{-\frac{s(\epsilon)\eta_\delta(\epsilon) N^{k\alpha}}{u^{\alpha k}}} du, \tag{73}
$$

where we redefine $\iota(\epsilon) \triangleq (1+\epsilon)^{k+1}(1-\epsilon)^{-(k+1)}$. Next, similarly to bounding $I_2(\delta)$ in (62 – 63), we use the change of variable $z = \iota(\epsilon)\eta_\delta(\epsilon)N^{\alpha k}u^{-\alpha k}$ to solve the integral in (73), we derive for N large

$$\mathbb{P}[C^{(N)} > \delta N] \geq (1-\epsilon)^3(\eta_\delta(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}}\frac{1-\alpha}{\alpha k}(\iota(\epsilon))^{\frac{1}{\alpha k}-\frac{1}{k}}\int_{\iota(\epsilon)\eta_\delta(\epsilon)}^{\left(\frac{N}{u_0}\right)^{k\alpha}\eta_\delta(\epsilon)\iota(\epsilon)} e^{-z}z^{\frac{1}{k}-\frac{1}{\alpha k}-1}dz,$$

which, after taking $\liminf_{N\to\infty}$ and letting $\epsilon \to 0$, renders

$$\liminf_{N\to\infty}\mathbb{P}[C^{(N)} > \delta N] \geq \frac{1-\alpha}{\alpha k}(\eta_\delta)^{\frac{1}{\alpha k}-\frac{1}{k}}\Gamma\left(\frac{1}{k}-\frac{1}{\alpha k},\eta_\delta\right),$$

where η_δ is the unique solution to the equation $f(\eta) = \delta$. The previous expression in conjunction with (65) concludes the proof of this theorem.

Finally, we prove the uniqueness of the solution η_δ for any $0 < \delta < 1$ and the limiting values for $F_k(\delta)$ when $k \to \infty$, $\delta \to 0$ and $\delta \to 1$. First, observe the function $f(\eta)$ as defined in (57). After expressing the function $\Gamma(-1/(\alpha k), \eta)$ in integral form and applying the integration by parts, we obtain

$$\begin{aligned} f(\eta) &= 1 - \frac{1}{\alpha k}\eta^{\frac{1}{\alpha k}}\int_\eta^\infty e^{-t}t^{-\frac{1}{\alpha k}-1}dt \\ &= 1 - e^{-\eta} + \eta^{\frac{1}{\alpha k}}\int_\eta^\infty e^{-t}t^{-\frac{1}{\alpha k}}dt. \end{aligned} \tag{74}$$

Now, since for any $\eta > 0$

$$\frac{d}{d\eta}f(\eta) = \frac{1}{\alpha k}\eta^{\frac{1}{\alpha k}-1}\int_\eta^\infty e^{-t}t^{-\frac{1}{\alpha k}}dt > 0,$$

we conclude that $f(\eta)$ is monotonically increasing in $\eta > 0$. Now, note that from (74), it is straightforward to conclude $f(\eta) > 0$ for any $\eta > 0$. Furthermore, since $f(\eta)$ is continuous function in $\eta \geq 0$ and

$$f(\eta) \leq 1 - e^{-\eta} + \eta^{\frac{1}{\alpha k}}\Gamma\left(1 - \frac{1}{\alpha k}\right) \to 0 \text{ as } \eta \to 0,$$

we conclude $\lim_{\eta\to 0} f(\eta) = 0$. Next, note that for large η, we can upper bound $f(\eta)$ as

$$f(\eta) = 1 - e^{-\eta} + \eta^{\frac{1}{\alpha k}}\int_\eta^\infty e^{-t}t^{-\frac{1}{\alpha k}}dt \leq 1 - e^{-\eta} + \eta^{\frac{1}{\alpha k}}\eta^{-\frac{1}{\alpha k}}e^{-\eta} = 1,$$

and, therefore

$$1 \geq f(\eta) \geq 1 - e^{-\eta} \to 1 \text{ as } \eta \to \infty.$$

Thus, $f(\eta)$ is strictly increasing continuous function for $\eta > 0$ with $\lim_{\eta\to 0} f(\eta) = 0$ and $\lim_{\eta\to\infty} f(\eta) = 1$. This implies that for any $0 < \delta < 1$, there is a unique solution $\eta(\delta) > 0$ of the equation $f(\eta) = \delta$.

Next, we prove the limitting value of the asymptotic expression in (49) when $k \to \infty$. Note that since

$$\Gamma\left(1-\frac{1}{\alpha k}\right) - \left(1-\frac{1}{\alpha k}\right)^{-1}\eta^{1-\frac{1}{\alpha k}} = \Gamma\left(1-\frac{1}{\alpha k}\right) - \int_0^\eta t^{-\frac{1}{\alpha k}}dt \leq \int_\eta^\infty e^{-t}t^{-\frac{1}{\alpha k}}dt \leq \Gamma\left(1-\frac{1}{\alpha k}\right), \tag{75}$$

then, using the continuity of $\Gamma(x)$ at $x = 1$, for any $\epsilon > 0$, there exists η_0 and k_0, such that for all $k \geq k_0$, $0 < \eta \leq \eta_0$

$$1 - \epsilon \leq \int_\eta^\infty e^{-t}t^{-\frac{1}{\alpha k}}dt \leq 1 + \epsilon. \tag{76}$$

Thus, for any $k \geq k_0$ and $\eta \leq \eta_0$

$$f(\eta) \geq f_1(\eta) \triangleq \eta^{\frac{1}{\alpha k}}(1-\epsilon). \tag{77}$$

Now, using (77) and the fact that functions $f(\eta)$, $f_1(\eta)$ are monotonically increasing, if the solution $\eta^*_{k,\delta}$ of the equation $f_1(\eta) = \delta$ satisfies $\eta^*_{k,\delta} \leq \eta_0$, it follows that the unique solution $\eta_{k,\delta}$ of the equation $f(\eta) = \delta$ must satisfy

$$\eta_{k,\delta} \leq \eta^*_{k,\delta}. \tag{78}$$

Then, from (77), it follows

$$\eta_{k,\delta} \leq \left(\frac{\delta}{1-\epsilon}\right)^{\alpha k},$$

and, therefore, if $\epsilon < 1-\delta$, there exists k_0 (possibly greater than before), such that for all $k \geq k_0$ inequality $\eta^*_{k,\delta} \leq \eta_0$ holds, which in conjunction with (78) yields

$$\lim_{k\to\infty} \eta_{k,\delta} = 0. \tag{79}$$

Next, it is not hard to check that

$$\begin{aligned}
\frac{d}{d\eta_{k,\delta}} F_k(\delta) &= \left(\frac{1-\alpha}{\alpha k}\right)^2 \eta_{k,\delta}^{\frac{1}{\alpha k}-\frac{1}{k}-1} \int_{\eta_{k,\delta}}^{\infty} e^{-t} t^{\frac{1}{k}-\frac{1}{\alpha k}-1} dt - \frac{1-\alpha}{\alpha k} e^{-\eta_{k,\delta}} \eta_{k,\delta}^{-1} \\
&\leq \left(\frac{1-\alpha}{\alpha k}\right)^2 \eta_{k,\delta}^{\frac{1}{\alpha k}-\frac{1}{k}-1} e^{-\eta_{k,\delta}} \int_{\eta_{k,\delta}}^{\infty} t^{\frac{1}{k}-\frac{1}{\alpha k}-1} dt - \frac{1-\alpha}{\alpha k} e^{-\eta_{k,\delta}} \eta_{k,\delta}^{-1} = 0,
\end{aligned}$$

and, therefore, for all k large, after using (78) and replacing $\eta_{k,\delta}$ with $\eta^*_{k,\delta}$ in expression (49), after applying integration by parts and similar arguments as in (75), we lower bound the tail of the search cost $F_k(\delta)$ as

$$\begin{aligned}
F_k(\delta) &\geq \frac{1-\alpha}{\alpha k} (\eta^*_{k,\delta})^{\frac{1}{\alpha k}-\frac{1}{k}} \int_{\eta^*_{k,\delta}}^{\infty} e^{-t} t^{\frac{1}{k}-\frac{1}{\alpha k}-1} dt \\
&= 1 - (\eta^*_{k,\delta})^{\frac{1}{\alpha k}-\frac{1}{k}} \Gamma\left(1 + \frac{1}{k} - \frac{1}{\alpha k}, \eta^*_{k,\delta}\right) \to 1 - \left(\frac{\delta}{1-\epsilon}\right)^{1-\alpha} \quad \text{as} \quad k \to \infty,
\end{aligned} \tag{80}$$

which, after letting $\epsilon \to 0$, implies

$$\lim_{k\to\infty} F_k(\delta) = 1 - \delta^{1-\alpha}. \tag{81}$$

Next, similarly as before, using the arguments that led to (76) and the limit in (79), for any $\epsilon > 0$ there exists $k_0 > 0$, such that for all $k \geq k_0$, inequality $\eta_{k,\delta} \leq \eta_0$ holds and, furthermore, we can upper bound $f(\eta)$ for all $\eta \leq \eta_0$ as

$$f(\eta) \leq f_2(\eta) \triangleq \epsilon + (1+\epsilon)\eta^{\frac{1}{\alpha k}}.$$

Then, using the analogous reasoning as before, for all $k \geq k_0$, the solution (say $\eta^*_{k,\delta} = ((\delta-\epsilon)/(1+\epsilon))^{\alpha k}$) to the equation $f_2(\eta) = \delta$ must be smaller than $\eta_{k,\delta}$. Thus, using the monotonicity of $F_k(\delta)$ in η_δ and similar arguments as in (80), after replacing η_δ with $\eta_{k,\delta}$ in (49), we upper bound the tail of the search cost and obtain

$$\lim_{k\to\infty} F_k(\delta) \leq 1 - \left(\frac{\delta-\epsilon}{1+\epsilon}\right)^{1-\alpha},$$

which, after letting $\epsilon \to 0$, in conjunction with (81), yields (50).

$\diamond$

The following theorem estimates the tail of the search cost distribution $\mathbb{P}[C^{(N)} > \delta N]$ as $N \to \infty$ in the case where $\alpha = 1$. Since the proof of this result uses completely analogous arguments to the ones used in the proof of Theorem 2, in order to avoid repetations we just state the result and present an outline of the proof.

Theorem 3 *Assume that $q_i = h_N/i$, $1 \le i \le N$, where h_N is the normalization constant. Then, for any $0 < \delta < 1$, as $N \to \infty$*

$$(\log N)\mathbb{P}[C^{(N)} > \delta N] \sim F_k(\delta) \triangleq \frac{1}{k}\Gamma(0, \eta_\delta), \tag{82}$$

where η_δ uniquely solves the equation

$$1 - \frac{1}{k}\eta^{\frac{1}{k}}\Gamma\left(-\frac{1}{k}, \eta\right) = \delta;$$

note that, $\Gamma(x, y)$, $y > 0$, is the incomplete Gamma function, i.e. $\Gamma(x, y) = \int_y^\infty e^{-t}t^{x-1}dt$. Furthermore, for any $0 < \delta < 1$,

$$\lim_{k\to\infty} F_k(\delta) = \log \delta. \tag{83}$$

Remarks: (i) In the context of the ordinary MTF searching on the list ($k = 1$), the convergence in distribution of the ratio $\log C^{(N)}/\log N$ to the uniform random variable on the unit interval is first proved in Lemma 4.7 of [4]. (ii) Similarly as in the remark (iii) after Theorem 2, it is possible to relax the assumption $q_i = h_N/i, 1 \le i \le N$. Again, by assuming that for any $\epsilon > 0$, there exists i_0 such that for all $i_0 \le i \le N$, inequality $(1-\epsilon)c/\log N \le q_i \le (1+\epsilon)c/\log N$ holds, the expression (84) would be replaced by this form, implying an almost identical formula to the one in (82), where the only difference is in the additional factor c on the right-hand-side.

Outline of the proof: In order to estimate the upper bound, we use the same arguments as in inequality (51). Note that for any $\epsilon > 0$ and N large enough, the normalization constant h_N is bounded by

$$(1-\epsilon)\frac{1}{\log N} < h_N < (1+\epsilon)\frac{1}{\log N}. \tag{84}$$

Thus, after bounding $\mathbb{E}S(t)$, similarly as in (52) and (54), assuming that t increases in N as $t = \eta(N\log N)^k/\xi(\epsilon)$, where $\eta > 0$ is a fixed constant, we obtain

$$\mathbb{E}S\left(\frac{\eta(N\log N)^k}{\xi(\epsilon)}\right) \le N - \int_{i_0}^{N} e^{-\frac{N^k\eta}{u^k}}du; \tag{85}$$

note that in this case $\xi(\epsilon)$ is defined as

$$\xi(\epsilon) \triangleq \frac{(1+\epsilon)^{k+1}\beta^{k-1}}{(k-1)!}.$$

Then, using the change of variable $z = N^k\eta u^{-k}$ the integral in (85) and applying the analogous steps as in (55 – 58), we obtain that for all $t \le t_0 \triangleq \eta_\delta(\epsilon_0)(N\log N)^k(\xi(\epsilon))^{-1}$, any $\epsilon_0 > 0$, N large and $\epsilon > 0$ small enough ($\epsilon < \epsilon_0\delta(1-2\epsilon_0)/(1-\delta(1-2\epsilon_0))$)

$$\mathbb{E}S(t) \le (1-\epsilon_0)\delta N,$$

where $\eta_\delta(\epsilon_0)$ is the unique solution to the equation

$$f(\eta) \triangleq 1 - \frac{1}{k}\eta^{\frac{1}{k}}\Gamma\left(-\frac{1}{k}, \eta\right) = \delta(1 - 2\epsilon_0).$$

Next, using the large deviation bound for the sum of N independent Bernoulli random variables, similarly as in (59), we derive, as $N \to \infty$,

$$\mathbb{P}[C^{(N)} > \delta N] \leq o\left(\frac{1}{\log N}\right) + \sum_{i=1}^{N} q_i \mathbb{P}[T_i \geq t_0]. \tag{86}$$

Now, after splitting the sum in the previous expression, analogously as in (60), we obtain, for N and i_0 large enough

$$\begin{aligned}\sum_{i=1}^{N} q_i \mathbb{P}[T_i \geq t_0] &\leq i_0 \mathbb{P}[T_{i_0} \geq t_0] + \sum_{i=i_0}^{N} q_i e^{-\frac{q_i^k \beta^{k-1}(1-\epsilon)}{(k-1)!} t_0} + \sum_{i=i_0}^{N} q_i e^{-h\epsilon q_i^{k-1} t_0} \\ &\triangleq I_1(\delta) + I_2(\delta) + I_3(\delta).\end{aligned} \tag{87}$$

Then, after replacing t_0 and using similar arguments that led to (61) and (64), we obtain

$$I_1(\delta) = o\left(\frac{1}{\log N}\right), \quad I_3(\delta) = o\left(\frac{1}{\log N}\right) \text{ as } N \to \infty. \tag{88}$$

Now, by upper bounding the sum in $I_2(\delta)$ with an integral, as in (62), applying the change of variable $z = (1-\epsilon)^{k+1}(1+\epsilon)^{-(k+1)}\eta_\delta(\epsilon_0)N^k u^{-k}$ and using the same arguments that led to (63), we conclude

$$I_2(\delta) \lesssim \frac{1}{\log N}\frac{1}{k}\Gamma(0, \eta_\delta) \text{ as } N \to \infty,$$

where η_δ uniquely solves the equation $f(\eta) = \delta$, and, therefore, in conjunction with (88), (87) and (86), yields the asymptotic upper bound for $\mathbb{P}[C > \delta N]$.

In order to prove the asymptotic lower bound for $\mathbb{P}[C > \delta N]$ in (82), we start by estimating $\mathbb{E}S(t)$ using the identical arguments as in the proof of the lower bound in Theorem 2. Thus, by letting t increase in N as $t = \eta(N \log N)^k/\xi(\epsilon)$, where $\eta > 0$ is a fixed constant, we define, similarly as in (66), for any $\epsilon > 0$ and i_0, N large enough

$$\begin{aligned}\mathbb{E}S\left(\frac{\eta(N\log N)^k}{\xi(\epsilon)}\right) &\geq N - i_0 - \int_{i_0}^{N} e^{-\frac{N^k \eta}{u^k}} du - \int_{i_0}^{N} e^{-h\epsilon \frac{N^k \log N}{u^{k-1}}} du \\ &\triangleq N - i_0 - I_1 - I_2,\end{aligned} \tag{89}$$

where $\xi(\epsilon)$ is redefined as $\xi(\epsilon) \triangleq (1-\epsilon)^{k+1}\beta^{k-1}((k-1)!)^{-1}$. Then, using the change of variable $z = N^k \eta u^{-k}$ for solving the integral of I_1 and the analogous arguments as in (68 – 71), we obtain that for all $t \geq t_0 \triangleq \eta_\delta(\epsilon)(N \log N)^k(\xi(\epsilon))^{-1}$, any $\epsilon > 0$ and N large, inequality

$$\mathbb{E}S(t) \geq (1+\epsilon)\delta N$$

holds, where $\eta_\delta(\epsilon)$ is the unique solution to the equation $f(\eta) = (1+2\epsilon)\delta$. Then, applying the same reasoning as in (72), in conjunction with the large deviation bound for the sum of independent Bernoulli random variables proved in Lemma 5 of the Appendix, we obtain that for N large

$$\mathbb{P}[C^{(N)} > \delta N] \geq (1-\epsilon)\sum_{i=1}^{N} q_i \mathbb{P}[T_i \geq t_0].$$

Next, by Lemma 3 and the analogous arguments as in (73), after using the change of variable $z = (1+\epsilon)^{k+1}(1-\epsilon)^{-(k+1)}\eta_\delta(\epsilon)N^k u^{-k}$ to solve the integral and letting $\epsilon \to 0$, we conclude

$$\mathbb{P}[C^{(N)} > \delta N] \gtrsim \frac{1}{\log N}\frac{1}{k}\Gamma(0,\eta_\delta) \text{ as } N \to \infty, \tag{90}$$

where η_δ is the unique solution to the equation $f(\eta) = \delta$. Thus, (90) in conjunction with the asymptotic upper bound from before proves (82).

Finally, it is left to prove the uniqueness of the solution η_δ of the equation $f(\eta) = \delta$ and the limit in (83). We omit the details of this proof since these properties follow directly from similar arguments as in the proof of Theorem 2.

$\diamond$

4 Numerical experiments

In this section to illustrate our main results stated in Theorems 1, 2 and 3, using simulation experiments. Since the asymptotic results are obtained for infinite number of documents N in Theorem 1, while in Theorems 2 and 3 the number N is passed to infinity, it can be expected that asymptotic expressions give reasonable approximation of the fault probability $\mathbb{P}[C^{(N)} > x]$, only if both N and x are large (with N much larger than x). However, our experiments show that the obtained approximations work well for relatively small values of N and almost all cache sizes $x < N$. Furthermore, our simulations validate significant improvement in performance of the introduced PAC(β, k), $k \geq 2$, algorithm when compared to the ordinary LRU scheme ($k = 1$), as predicted by our asymptotic results.

4.1 Convergence to stationarity

In order to ensure that the simulated values of the fault probabilities do not deviate significantly from the stationary ones, we first estimate the difference between the distributions of $C^{(N)}$ and $C_n^{(N)}$, where $C_n^{(N)}$ is the search cost after n requests with arbitrary initila conditions. Thus, using (2 – 4) with $\epsilon = 1/2$, we upper bound the difference between the tails of these distributions as

$$\sup_x \left|\mathbb{P}[C_n^{(N)} > x] - \mathbb{P}[C^{(N)} > x]\right| \leq e_n \triangleq \mathbb{P}\left[\tau_n < \frac{n}{2}\right] + \sum_{i=1}^{N} q_i \mathbb{P}\left[T_i > \frac{n}{2} - \beta\right],$$

where τ_n is the nth arrival point in a Poisson process of unit rate. Thus, by applying the bound (14) and setting $\epsilon = 1/2$, yields

$$e_n \leq \mathbb{P}\left[\tau_n < \frac{n}{2}\right] + \sum_{i=1}^{N} q_i \left[e^{-p_i q_i \frac{\beta}{4}(\frac{n}{2}-\beta)} + (1-p_i)^{\frac{\frac{n}{2}-\beta}{4\beta}-1}\right]; \tag{91}$$

recall that probabilities $p_i = \mathbb{P}[M_\beta^{(q_i)} \geq k-1]$, $1 \leq i \leq N$, where $M_t^{(q)}$ is a counting Poisson process of rate q. The first term in expression (91) is easy to estimate since $\mathbb{P}[\tau_n < n/2] = \mathbb{P}[M_{n/2}^{(1)} > n]$; in addition, since the Poisson distribution is highly concentrated around the mean, this term converges very fast to zero. Therefore, it is easy to see that the error bound in (91) is dominated by the sum. Furthermore, the value of the sum decreases as β increases since $p_i = O(\beta^{k-1}q_i^{k-1})$. Hence, the increase of the parameter β

speeds up the convergence of the search cost process $\{C_n^{(N)}\}$ to stationarity. This makes the algorithm more adaptable to the possible fluctuations in the document popularities. On the other hand, the larger β implies the larger expected size of the additional storage needed to keep track of the past requests. Thus, although the stationary performance of the PAC algorithm is invariant to β, this parameter provides an important design component whose choice has to balance between the algorithm complexity and adaptability.

4.2 Experiments

In the presented experiments we take the number of documents to be $N = 1300$ with popularities satisfying $q_i = h_N/i^\alpha$, $1 \le i \le 1300$, where $h_N = (\sum_{i=1}^N 1/i^\alpha)^{-1}$. Also, we select $\beta = 20$ and α : (1) $\alpha = 1.2$, (2) $\alpha = 1$ and (3) $\alpha = 0.8$. In each experiment, before we conduct measurements, we allow the time of the first $n = 10^{10}$ requests to be a warm-up time for the system to reach stationarity. Then, the actual measurement time is also set to be 10^{10} requests long. The initial permutation of the list is chosen uniformly at random and the initial set of requests in $(-\beta, 0)$ is taken to be empty. The fault probabilities are measured for cache sizes $x = 50j$, $1 \le j \le 15$. Simulation results are presented with "*" symbols on Figures 3, 4 and 5, while our approximations are presented with the solid lines on the same figures.

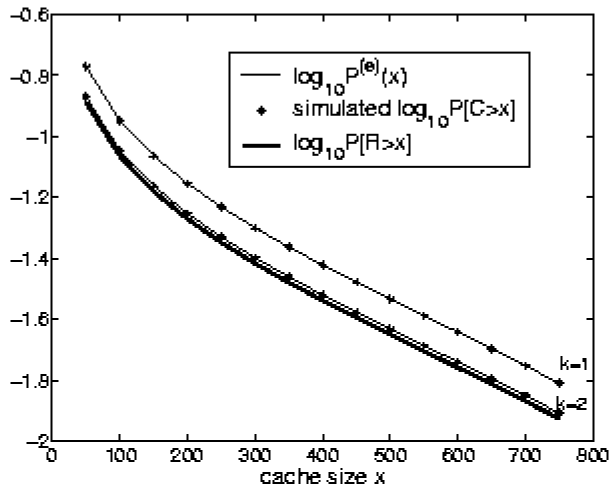

Figure 3: Illustration for the case of $\alpha = 1.2$

Experiment 1 We set $\alpha = 1.2$ and measure the cache fault probabilities for values PAC$(20, k)$, $k = 1, 2$, algorithm. We compare the obtained measurements with our approximation given by $\mathbb{P}^{(e)}(x) = K_k(\alpha)\mathbb{P}[R > x]$, as implied by Theorem 1. The experimental results for the cases when $k \ge 3$ are almost indistinguishable from the performance of the optimal algorithm, $\mathbb{P}[R > x]$, and for that reason we did not present them on Figure 3. After estimating e_n in (91) for a given warm-up time from above, we obtain that $e_n < 10^{-6}$ for both $k = 1, 2$, which is negligible when compared to the smallest measured probabilities ($> 10^{-2}$) and, therefore, the measured fault probabilities are essentially the stationary ones. Figure 3 shows an excellent agreement between the approximation $P^{(e)}(x)$ and experimental results, as well as a significant improvement in performance for $k = 2$.

Experiment 2 Here, we select $\alpha = 1$ and measure the cache fault probabilities for $k = 1, 2, 3$. Since the normalization constant $h_N = \log N + \gamma + o(1)$ as $N \to \infty$, where γ is the Euler's constant, the ratio $h_N/\log N$ converges slowly to one and, therefore, instead of using the approximation $\mathbb{P}[C^{(N)} > x] \approx (\log N)F_k(x/N)$, as suggested by Theorem 3, we define $\mathbb{P}^{(e)}(x) = h_N F_k(x/N)$. Furthermore, after

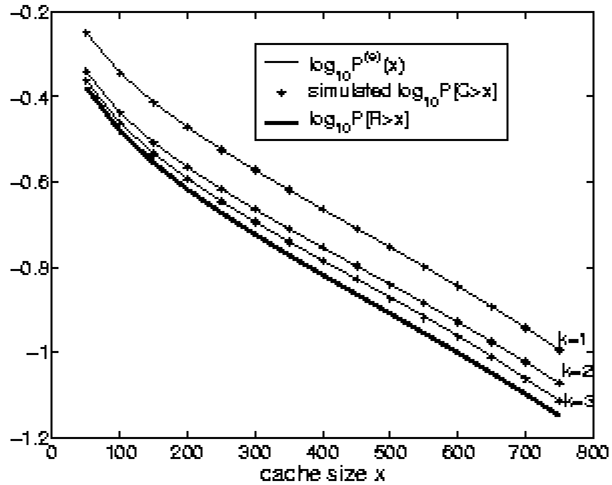

Figure 4: Illustration for the case of $\alpha = 1$

computing e_n, we obtain that for $k = 1, 2$, the error $e_n < 10^{-6}$, while for $k = 3$ it satisfies $e_n < 0.004$, which is insignificant when compared to the smallest values of the measured probabilities and, thus, the measured process is basically in stationarity. Again, the accuracy of the approximation $P^{(e)}(x)$ and the improvement in performance is apparent from Figure 4.

Experiment 3 Finally, the third example assumes $\alpha = 0.8$ and considers cases $k = 1, 2, 3$. Similarly as in the case of $\alpha = 1$, due to the slow convergence of $h_N N^{1-\alpha}/(1-\alpha)$ to one as $N \to \infty$, we use an estimate $\mathbb{P}^{(e)}(x) = h_N(N^{1-\alpha}/(1-\alpha))F_k(x/N)$ instead of $F_k(x/N)$ that can be inferred from Theorem 2. After estimating the error e_n, we obtain that for all $k = 1, 2, 3$ it satisfies $e_n < 10^{-6}$, which again is much smaller than the smallest measured value of the fault probability and hence, the process is effectively in stationarity. Once again, the validity of the approximation $P^{(e)}(x)$ and the benefit of the PAC algorithm is evident from Figure 5.

5 Concluding remarks

In this paper we propose a new LRU-based, PAC(β, k) replacement rule that possesses all of the desirable properties of the LRU policy, such as law complexity, ease of implementation, adaptability to variability in access patterns. In the case of independent reference model, we derive that the performance (fault probability) of the PAC policy, for large cache sizes, can approach arbitrarily close to the optimal, frequency algorithm, with a negligible additional complexity. Our model considers request process satisfying generalized Zipf's law popularity distribution and arriving at Poisson time points. Our analytical approach uses probabilistic (average-case) analysis that relies on the well-known connection between the LRU caching and Move-to-Front searching scheme. We exploit the novel large deviation approach introduced recently in [11]. Our results are further validated using simulation. Note that β does not influence the asymptotic performance, but still, given the observations in Subsection 4.1, it is an important design parameter representing the tradeoff between the faster adaptability (small β) and shorter transience (larger β). Furthermore, simulations show a significant improvement of the PAC algorithm in comparison to the ordinary LRU scheme, even for a small values of k ($k = 2, 3$). The excellent agreement between the analytical and experimental results and simplicity of the PAC algorithm itself, implies a very high potential of the proposed policy for practical purposes.

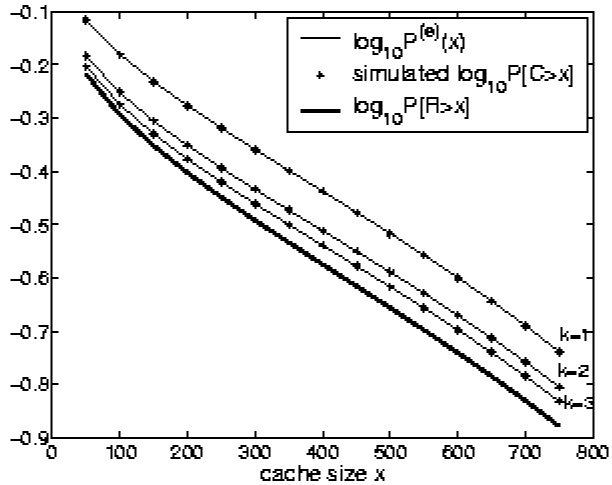

Figure 5: Illustration for the case of $\alpha = 0.8$

Given the analytic approach established in our recent work on the analysis of the LRU policy in the presence of the dependent request traffic [11] and investigation of the randomized LRU policies that consider documents with different sizes [10], we conjecture that analogous results hold for the PAC algorithm as well. In the context of the ordinary MTF with lighter tailed request distributions, the asymptotic analysis on the fluid limit of the search cost derived in [8] could be analogously applied to the case of the PAC policy as well. Finally, we would like to note that our algorithm relates to the earlier proposed "k-in-a-row" rule [12, 7]. This rule was stated in the context of the expected list search cost. However, its analytical performance in the caching framework remains unknown.

Appendix

Lemma 4 *Let $S(t) = \sum_{i=1}^{\infty} B_i(t)$ and assume $q_i \sim c/i^{\alpha}$ as $i \to \infty$, with $\alpha > 1$ and $c > 0$. Then, as $t \to \infty$,*

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

This lemma is originally stated and proved in [8].

Lemma 5 *Let $\{B_i, 1 \leq i \leq N\}$, $N \leq \infty$, be a sequence of independent Bernoulli random variables, $S = \sum_{i=1}^{N} B_i$ and $m = \mathbb{E}[S]$. Then for any $\epsilon > 0$, there exists $\theta_\epsilon > 0$, such that*

$$\mathbb{P}[|S - m| > m\epsilon] \leq 2e^{-\theta_\epsilon m}.$$

This lemma, in slightly different form, is stated and proved in [11].

References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, 1974.

[2] 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.

[3] E. Cinlar. *Introduction to Stochastic Processes*. Prentice–Hall, 1975.

[4] 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.

[5] J. A. Fill and L. Holst. On the distribution of search cost for the move-to-front rule. *Random Structures and Algorithms*, 8(3):179, 1996.

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

[7] G. H. Gonnet, J. I. Munro, and H. Suwanda. Exegesis of self-organizing linear search. *SIAM J. Comput.*, 10(3):613–637, 1981.

[8] 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.

[9] P. R. Jelenković. Least-Recently-Used Caching with Zipf's Law Requests. In *The Sixth INFORMS Telecommunications Conference*, Boca Raton, Florida, 2002.

[10] P. R. Jelenković and A. Radovanović. Optimizing LRU caching for variable document sizes. Combinatorics, Probability & Computing, to appear., 2002.

[11] P. R. Jelenković and A. Radovanović. Least-recently-used caching with dependent requests. Theoretical computer science, to appear., 2003.

[12] Y. C. Kan and S. M. Ross. Optimal list order under partial memory constraints. *Journal of Applied Probability*, 17:1004–1015, 1980.

[13] Tracy Kimbrel. Online paging and file caching with expiration times. *Theoretical Computer Science*, 268:119–131, 2001.

[14] Elizabeth J. O'Neil, Patrick E. O'Neil, and Gerhard Weikum. An Optimality Proof of the LRU-K Page Replacement Algorithm. *Journal of the ACM*, 46:92–112, 1999.