LRU Cache under Stationary Requests

UMass CICS Technical Report UM-CS-2017-005

Bo Jiang*, Philippe Nain**, and Don Towsley*

*University of Massachusetts Amherst
**Inria, Grenoble - Rhône-Alpes
{bjiang, towsley}@cs.umass.edu, philippe.nain@inria.fr

Abstract

In this paper we focus on the LRU cache under the independence reference model to systems where requests for different contents are described by independent stationary and ergodic processes. We extend a TTL-based approximation of the cache hit rate, first proposed by Fagin, [7], for the independence reference model to this more general workload model. We further show that this approximation becomes exact as the number of contents goes to infinity while the ratio of number of contents to cache size remains constant. Moreover we establish this not only for the aggregate cache hit rate but for every individual content. Last, we obtain the rate of convergence.

1 Introduction

Caches form a key component of many computer networks and systems. Moreover, they are becoming increasingly more important with the current development of new content-centric network architectures. A variety of cache replacement algorithms has been introduced and analyzed over the last few decades, most based on the least recently used algorithm (LRU). Considerable work has focused on analyzing these policies [2, 3, 8, 15] for iid requests (the so-called independent reference model - IRM) and for Markov-modulated requests [6, 13, 14].

However, with the exception of time-to-live (TTL) caches [5], networks of caches defy exact analysis and only approximations have been developed [4, 17]. The link between an LRU cache and a TTL cache has been first pointed out in [7]. In this paper, Fagin introduced the concept of a characteristic time (our terminology) and showed asymptotically that the performance of LRU converges to that of a TTL cache with a timer set to the characteristic time. With the exception of an application to caching in [8], this work disappeared and several papers [4, 10, 12] reintroduced the approximation for LRU, its variants, and other cache policies (FIFO, random). More recently, [11] extended the characteristic time (CCT) approximation to a setting where requests for distinct contents are independent and described by renewal processes. The accuracy of this approximation is supported by simulations but a theoretical basis is lacking. Providing a theoretical justification of this extended CCT approximation is the focus of this paper.

The main contribution of this paper is an extension of Fagin's results for LRU under IRM assumptions to the more general setting where requests for different content are independent of each other but requests to each content are described by a stationary and ergodic process. Based on these results, we develop a CCT approximation for the performance of an LRU cache. Furthermore, we provide simple closed form bound on the approximation error, which yields the rate of convergence of the approximation to the asymptotic limit as the cache size and number of contents increase to infinity while keeping the ratio of the two constant. Our approach is similar to that of [16], which provides an error bound for CCT approximation under shot noise request model. The bound in [16] is in the form of an optimization problem and requires numerical computation. For stationary and ergodic request processes, we are able to bound the value of the corresponding optimization problem analytically under some mild conditions.

The rest of the paper is organized as follows. Section 2 presents our model of an LRU cache under a general request model. Section 3 presents the extension of Fagin's result to the case where requests for contents are described by independent stationary and ergodic processes. We then establish convergence results for all individual contents in Section 4 and convergence rates in Section 5. Section 6 extends the results to cover scenarios where different content providers with different request workloads share an LRU cache. Last concluding statements are provided in Section 7.

2 Model

We consider a cache of size C serving n unit size contents labelled $i = 1, \ldots, n$ where $C \in (0, n)$. Requests for the contents are described by n independent stationary and ergodic point processes $N_i := \{t_i(k), k \in \mathbb{Z}\}$, where $-\infty \leq \cdots < t_i(-1) < t_i(0) \leq 0 < t_i(1) < \cdots \leq \infty$ represent the successive request times to content $i = 1, \ldots, n$ having probability measure $\mathbb{P}$ and associated expectation operator $\mathbb{E}$. Let $0 < \lambda_i < \infty$ denote the intensity of request process N_i, i.e., the long term average request rate for content i (see e.g. [1, Sections 1.1 and 1.6] for an introduction to stationary and ergodic point processes).

Let $\mathbb{P}_i^0$ be the Palm probability associated with the point process N_i (see e.g. [1, Eq. (1.2.1)]). In particular, $\mathbb{P}_i^0(\{t_i(0) = 0\}) = 1$. In other words, under $\mathbb{P}_i^0$ content i is requested at time $t = 0$. Define $G_i(x) = \mathbb{P}_i^0(t_i(1) \leq x)$, the cdf of the duration between two successive requests to content i under $\mathbb{P}_i^0$. It is known that $\mathbb{E}_i^0[t_i(1)] = 1/\lambda_i$ [1, Exercice 1.2.1], with $\mathbb{E}_i^0$ the expectation operator associated with $\mathbb{P}_i^0$.

Last, we define $\mathbb{P}^0$, the Palm probability associated with the point process $\{t(k), k \in \mathbb{Z}\}$, $-\infty \leq \cdots < t(-1) < t(0) \leq 0 < t(1) < \cdots \leq \infty$, resulting from the superposition of the n independent point processes $N_1, \ldots, N_n$. Under $\mathbb{P}^0$ a content is requested at $t = 0$ (since $\mathbb{P}^0[\{t(0) = 0\}] = 1$). Let $X_0 \in \{1, \ldots, n\}$ denote this content. We denote by $\mathbb{E}^0$ the expectation operator associated with $\mathbb{P}^0$. It is known that (see e.g. [1, Section 1.4.2])

$$\mathbb{P}^0[\{t_i(0) = i\}] = \frac{\lambda_i}{\Lambda^{(n)}} := p_i^{(n)}, \tag{1}$$

with $\Lambda^{(n)} := \sum_{i=1}^n \lambda_i$. We assume that $\Lambda := \lim_{n\to\infty} \Lambda^{(n)} \in (0, \infty)$.

For any cdf F with support in $[0, \infty)$, let

$$\hat{F}(x) = \frac{1}{\mathbb{E}F} \int_0^x \bar{F}(y)dy, \tag{2}$$

where $\bar{F} = 1 - F$ is the ccdf, and $\mathbb{E}F = \int_0^\infty \bar{F}(y)dy$ is the mean. It is well-known that (see e.g. [1, Section 1.3.4])

$$\mathbb{P}[-t_i(0) \leq x] = \mathbb{P}[t_i(1) \leq x] = \lambda_i \int_0^x \bar{G}_i(y)dy = \hat{G}_i(x) \tag{3}$$

for each i. When moving forward (resp. backward) in time, $\mathbb{P}[-t_i(0) \leq x]$ and $\mathbb{P}[t_i(1) \leq x]$ are the cdfs of the age (resp. residual time) and residual time (resp. age), respectively, associated with the inter-request times of content i.

We assume that

$$G_i(x) = G(\lambda_i x), \tag{4}$$

for some cdf G with mean 1. Note that (4) holds if $G_i(\cdot)$ is the exponential distribution. It follows from (2) and (4) that

$$\hat{G}_i(x) = \hat{G}(\lambda_i x). \tag{5}$$

We also assume that there exists a continuously differentiable cdf F with support in $[0, 1]$ such that for $i = 1, 2, \ldots, n$,

$$p_i^{(n)} = F\left(\frac{i}{n}\right) - F\left(\frac{i-1}{n}\right) = \frac{1}{n} F'(\xi_i^{(n)}), \tag{6}$$

where $\xi_i^{(n)} \in \left(\frac{i-1}{n}, \frac{i}{n}\right)$. The existence of $\xi_i^{(n)}$ is guaranteed by the mean-value theorem. We assume that $F'(x) > 0$ a.e. on $[0,1]$. We allow $F'(0)$ to be infinite, to allow Zipf's law in particular.

Let $Y_i(t) = 1$ if content i was requested during the interval $[-t, 0)$ and $Y_i(t) = 0$ otherwise. With this notation, $Y(t) := \sum_{i=1}^n Y_i(t)$ is the number of distinct contents requested during $[-t, 0)$. Let $[-\tau, 0)$ be the smallest past interval such that there have been C distinct contents referenced in that interval, i.e.,

$$\tau = \inf\{t : Y(t) \geq C\}.$$

Note that if we reverse the arrow of time, we obtain in steady-state statistically the same request processes, and τ is a stopping time for the process $Y(t)$. The stationary hit probability of an LRU cache is then given by

$$H^{\mathrm{LRU}} = \mathbb{P}^0[Y_{X_0}(\tau) = 1]. \tag{7}$$

If the cache is a TTL cache with timer T, the hit probability is

$$H^{\mathrm{TTL}}(T) = \mathbb{P}^0[Y_{X_0}(T) = 1]. \tag{8}$$

More specifically, $\mathbb{P}_i^0[Y_i(\tau) = 1]$ and $\mathbb{P}_i^0[Y_i(T) = 1]$ are the stationary hit probabilities of content i in an LRU cache and in a TTL cache with timer T, respectively. Observe that

$$H^{\mathrm{LRU}} = \sum_{i=1}^n p_i^{(n)} \mathbb{P}_i^0[Y_i(\tau) = 1] \quad \text{and} \quad H^{\mathrm{TTL}}(T) = \sum_{i=1}^n p_i^{(n)} \mathbb{P}_i^0[Y_i(T) = 1]. \tag{9}$$

Define

$$\beta^\star(\nu) = \int_0^1 \hat{G}(\nu F'(x))dx \quad \text{and} \quad h^\star(\nu) = \int_0^1 F'(x) G(\nu F'(x))dx. \tag{10}$$

In the next section, we show that, as n becomes large, an LRU cache behaves as a TTL cache with a timer value that we identify.

3 Asymptotic behavior

Throughout $T_n(\nu) = n\nu/\Lambda^{(n)}$. This section is devoted to the proof of the following result:

Proposition 3.1. *Assume that $C \sim n\beta_0$ with $\beta_0 > 0$. Then,*

$$\begin{aligned} \lim_{n\to\infty} H^{\mathrm{LRU}} &= \lim_{n\to\infty} H^{\mathrm{TTL}}(T_n(\nu_0)) && (11) \\ &= h^\star(\nu_0), && (12) \end{aligned}$$

where ν_0 is the unique solution in $(0, \infty)$ of $\beta^\star(\nu) = \beta_0$.

This result shows that, when the number of contents n is large, the hit probability of a LRU cache of size $n\beta_0$ is close to the hit probability of a TTL cache with timer $T \sim \nu_0 n/\Lambda^{(n)}$. This result was first proved rigorously by Fagin [7] in the IRM setting. Our result provides a rigorous extension of Fagin's result to the case when successive requests to each content follow a stationary and ergodic process and when these content request processes are mutually independent.

Fagin's work went mostly unnoticed (although cited in [8]) and the connexion between LRU and TTL caches in the IRM setting was rediscovered in [4] through simple but non-rigorous arguments. On the other hand, [4] was one of the first attempts (with [17] and later [5]) to study a network of caches and to develop approximations for the performance metrics of interest (hit rate, etc.) by using the connexion between LRU and TTL caches.

Proposition 3.1 holds for a unique content popularity cdf (see (6)). Its extension to several content popularity probability distributions is addressed in Section 6. Such a model may be useful when, for instance, several content providers share a common LRU cache and contents associated with different providers exhibit different popularity probability distributions.

The proof of Proposition 3.1 relies on the eight lemmas stated and proved below.

Lemma 3.2. *For $i, j = 1, \ldots, n$, $t > 0$,*

$$\mathbb{P}^0_j[Y_i(t) = 1] = \mathbb{1}_{\{j=i\}} G_i(t) + \mathbb{1}_{\{j \neq i\}} \hat{G}_i(t). \tag{13}$$

Furthermore, $Y_1(t), \ldots, Y_n(t)$ are mutually independent given X_0.

Proof. Assume first that $j = i$. Then,

$$\mathbb{P}^0_i[Y_i(t) = 1] = \mathbb{P}^0_i[-t_i(-1) \leq t] = G_i(t).$$

Assume now that $j \neq i$. We have

$$\mathbb{P}^0_j[Y_i(t) = 1] = \mathbb{P}[-t_i(0) \leq t] = \hat{G}_i(t)$$

from (3). This proves (13).

The second statement of the lemma is a consequence of the independence of the point processes $\{t_1(k), k \in \mathbb{Z}\}, \ldots, \{t_n(k), k \in \mathbb{Z}\}$. □

For a TTL cache with timer T define

- $C(T)$ as the expected number of contents in the cache;
- $C^0_i(T)$ as the expected number of contents in the cache seen by a request for content i

Note that

$$C(T) = \mathbb{E}[Y(T)] \quad \text{and} \quad C^0_i(T) = \mathbb{E}^0_i[Y(T)]. \tag{14}$$

Lemma 3.3.

$$\begin{aligned} C(T) &= \sum_{i=1}^n \hat{G}_i(T) & (15) \\ C^0_i(T) &= C(T) + G_i(T) - \hat{G}_i(T). & (16) \end{aligned}$$

The mapping $T \to C(T)$ concave. Furthermore, it is strictly increasing for all T such that $C(T) < n$.

Proof. We have

$$C(T) = \mathbb{E}[Y(T)] = \sum_{i=1}^n \mathbb{P}[Y_i(T) = 1] = \sum_{i=1}^n \mathbb{P}[-t_i(0) \leq T] = \sum_{i=1}^n \hat{G}_i(T), \tag{17}$$

from (3), which proves (15). To prove (16) observe that

$$C^0_i(T) = \mathbb{E}^0_i[Y(T)] = \sum_{j=1}^n \mathbb{P}^0_i[Y_j(T) = 1] = \sum_{\substack{j=1 \\ j \neq i}}^n \hat{G}_j(T) + G_i(T) = C(T) + G_i(T) - \hat{G}_i(T),$$

by using (13) and (15).

Note that $C'(T) = \sum_{i=1}^n \lambda_i(1 - G_i(T))$. Since $C'(T)$ is a decreasing function of T, it follows that $C(T)$ is concave.

Since $C'(T) \geq 0$, we conclude that $C(T)$ is non-decreasing. Assume that $C'(T) = 0$ for some $T > 0$. Then, $G_i(T) = 1$ for each i which yields $\hat{G}_i(T) = 1$ for each i (Hint: $G_i(t) = 1$ for all $t \geq T$) which in turn implies that $C(T) = n$. Therefore, $C'(T) > 0$ for all T such that $C(T) < n$, which proves that the mapping $T \mapsto C(T)$ is strictly increasing at such T. □

For a TTL cache with timer $T_n(\nu)$ define

- $\beta^{(n)}(\nu) = (1/n)C(T_n(\nu))$, the expected fraction of contents in the cache;

- $\beta_i^{0,(n)}(\nu) = (1/n)C_i^0(T_n(\nu))$, the expected fraction of contents in the cache seen by a request for content i.

The next two lemmas investigate the limiting behavior of $\beta^{(n)}(\nu)$ and $\beta_i^{0,(n)}(\nu)$ as $n \to \infty$.

Lemma 3.4.

$$\lim_{n\to\infty} \beta^{(n)}(\nu) = \beta^\star(\nu), \tag{18}$$

where $\beta^\star(\nu)$ is defined in (10). Moreover,

$$\frac{d}{d\nu}\beta^\star(\nu) = 1 - h^\star(\nu) \geq 0, \tag{19}$$

which is strictly positive for all $\nu \geq 0$ such that $\beta^\star(\nu) < 1$.

Proof. With (15) we get

$$\begin{aligned}\beta^{(n)}(\nu) &= \frac{1}{n}\sum_{i=1}^n \hat{G}_i(T_n(\nu)) = \frac{1}{n}\sum_{i=1}^n \hat{G}(p_i^{(n)}\Lambda^{(n)}T_n(\nu)) \quad \text{from (5)}\\ &= \frac{1}{n}\sum_{i=1}^n \hat{G}(\nu F'(\xi_i^{(n)})) \quad \text{from (6)}\\ &\to \int_0^1 \hat{G}(\nu F'(x))dx = \beta^\star(\nu) \quad \text{as } n \to \infty.\end{aligned}$$

Differentiating $h^\star(\nu)$ in (10) wrt to ν readily gives the rhs of (19). Note that we have pushed differentiation inside the integral sign by Theorem 2.27 of [9]. If $\frac{d}{d\nu}\beta^\star(\nu) = 0$, then $h^\star(\nu) = 1$. Since $h^\star(\nu) \leq 1$ and $F'(x) > 0$ a.e. on $[0,1]$, we must have $G(\nu F'(x)) = 1$ a.e. on $[0,1]$, in which case

$$\hat{G}(\nu F'(x)) = \int_0^{\nu F'(x)} \bar{G}(y)dy = \int_0^\infty \bar{G}(y)dy = 1$$

a.e. on $[0,1]$, which implies $\beta^\star(\nu) = 1$. This concludes the proof. □

Lemma 3.5. *For $i = 1,\ldots,n$ and $\nu > 0$,*

$$\left|\beta_i^{0,(n)}(\nu) - \beta^{(n)}(\nu)\right| \leq \frac{1}{n}.$$

In particular, $\lim_{n\to\infty} \beta_i^{0,(n)}(\nu) = \lim_{n\to\infty} \beta^{(n)}(\nu) = \beta^\star(\nu)$ for all $i = 1,\ldots,n$ and $\nu > 0$.

Proof. By Lemma 3.3

$$|\beta_i^{0,(n)}(\nu) - \beta^{(n)}(\nu)| = \frac{1}{n}|G_i(Y_j(T_n(\nu))) - \hat{G}_i(Y_j(T_n(\nu)))| \leq \frac{1}{n}$$

by using the fact that G_i and $\hat{G}_i$ are both cdfs, and the lemma follows. □

Note that if request processes were Poisson processes the identity $\beta^{(n)}(\nu) = \beta^{0,(n)}(\nu)$ would hold since $G_i(\cdot) = \hat{G}_i(\cdot)$ for each i. A result which would also follow from the PASTA property.

The next lemmas focus on the hit probability in a TTL cache with timer T both for finite n and when $n \to \infty$.

Lemma 3.6.

$$H^{\text{TTL}}(T) = \sum_{i=1}^n p_i^{(n)} G_i(T).$$

Proof. From (8) we obtain

$$H^{\mathrm{TTL}}(T) = \sum_{i=1}^{n} p_i^{(n)} \mathbb{P}_i^0[Y_i(T) = 1] = \sum_{i=1}^{n} p_i^{(n)} G_i(T).$$

□

Lemma 3.7.

$$\lim_{n\to\infty} H^{TTL}(T_n(\nu)) = h^\star(\nu),$$

where $h^\star(\nu)$ is defined in (10).

Proof. From Lemma 3.6

$$\begin{aligned} H^{TTL}(T_n(\nu)) &= \sum_{i=1}^{n} p_i^{(n)} G_i(T_n(\nu)) \\ &= \sum_{i=1}^{n} p_i^{(n)} G(\lambda_i T_n(\nu)) \quad \text{from (5)} \\ &= \sum_{i=1}^{n} \frac{1}{n} F'(\xi_i^{(n)}) G(\nu F'(\xi_i^{(n)})) \quad \text{from (6)} \\ &\to \int_0^1 F'(x) G(\nu F'(x)) dx = h^\star(\nu) \quad \text{as } n\to\infty. \end{aligned}$$

□

Lemma 3.8.

$$\mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \le T] \le \mathbb{P}^0[Y_{X_0}(T) = 1, \tau \le T],$$

and

$$\mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \ge T] \ge \mathbb{P}^0[Y_{X_0}(T) = 1, \tau \ge T].$$

Proof. Since $Y_{X_0}(t)$ is increasing in t, the inequalities follow from a simple sample path argument. □

The next lemma shows that τ is concentrated around $T^{(n)}(\nu_0)$. It is given in a form that is more general than needed in this section, but we will need this form in Section 4.

Lemma 3.9. *Assume that $C \sim \beta_0 n$ with $\beta_0 \in (0,1)$ and let ν_0 be the unique solution of $\beta^\star(\nu) = \beta_0$. For $\nu_1 < \nu_0 < \nu_2$, as $n \to \infty$,*

$$\max_{1\le i\le n} \mathbb{P}_i^0[\tau < T_{\nu_1}^{(n)}] \to 0,$$

and

$$\max_{1\le i\le n} \mathbb{P}_i^0[\tau > T_{\nu_2}^{(n)}] \to 0.$$

Proof. Since $\beta^\star(0) = 0$, $\lim_{\nu\to\infty} \beta^\star(\nu) = 1$ and the mapping $\nu \mapsto \beta^\star(\nu)$ is continuous, the equation $\beta^\star(\nu) = \beta_0$ has at least one solution in $(0,\infty)$ when $\beta_0 \in (0,1)$. The uniqueness of this solution comes from Lemma 3.4.

Thanks to Lemma 3.5 we have

$$\frac{C-1}{n} - \beta^{(n)}(\nu) \le \frac{C}{n} - \beta_i^{0,(n)}(\nu) \le \frac{C+1}{n} - \beta^{(n)}(\nu) \tag{20}$$

for $i = 1, \ldots, n$.

Since $\beta^\star(\nu_0) = \beta_0 < 1$, we know by Lemma 3.4 that $d\beta^\star(\nu)/d\nu > 0$ at $\nu = \nu_0$ so that $\beta_0 - \beta^\star(\nu_1) > 0$ for all $0 < \nu_1 < \nu_0$ and $\beta_0 - \beta^\star(\nu_2) < 0$ for all $\nu_2 > \nu_0$ by using the fact that the mapping $\nu \mapsto \beta(\nu)$ is non-decreasing. Consequently, for $0 < \nu_1 < \nu_0$,

$$\lim_{n\to\infty} \left(\frac{C-1}{n} - \beta^{(n)}(\nu_1) \right) = \beta_0 - \beta^\star(\nu_1) > 0, \tag{21}$$

and for $\nu_2 > \nu_0$,

$$\lim_{n\to\infty}\left(\frac{C+1}{n} - \beta^{(n)}(\nu_2)\right) = \beta_0 - \beta^{\star}(\nu_2) < 0, \tag{22}$$

by using Lemma 18.

Recall the definition of τ,

$$\tau = \inf\{t : Y(t) \geq C\}.$$

Fix $0 < \nu_1 < \nu_0$. By (21) we know that there exists N_1 such that $(C-1)/n - \beta^{(n)}(\nu_1) > 0$ for $n > N_1$. Hence, for $n > N_1$,

$$\begin{aligned}
\mathbb{P}_i^0[\tau < T_n(\nu_1)] &\leq \mathbb{P}_i^0[Y(T_n(\nu_1)) \geq C] \\
&= \mathbb{P}_i^0[Y(T_n(\nu_1)) - \mathbb{E}_i^0[Y(T_n(\nu_1))] \geq C - n\beta_i^{0,(n)}(\nu_1)] \text{ as } \mathbb{E}_i^0[Y(T_n(\nu_1))] = n\beta_i^{0,(n)}(\nu_1) \\
&\leq P_i^0[Y(T_n(\nu_1)) - \mathbb{E}_i^0[Y(T_n(\nu_1))] \geq n((C-1)/n - \beta^{(n)}(\nu_1))] \text{ from (20)} \\
&\leq e^{-2n^{-1}(C-1-n\beta^{(n)}(\nu_1))^2} = e^{-2n((C-1)/n-\beta^{(n)}(\nu_1))^2} \qquad (23)
\end{aligned}$$

for $i = 1, \ldots, n$, where (23) holds from Hoeffding's inequality. Therefore, as $n \to \infty$,

$$\max_{1\leq i\leq n} \mathbb{P}_i^0[\tau < T_{\nu_1}^{(n)}] \leq \exp\left\{-2n\left[\frac{C-1}{n} - \beta^{(n)}(\nu_1)\right]^2\right\} \to 0.$$

Fix $\nu < \nu_2$. By (22) we know that there exists N_2 such that $(C+1)/n - \beta^{(n)}(\nu_2) < 0$ for $n > N_2$. Hence, for $n > N_2$,

$$\begin{aligned}
\mathbb{P}_i^0[\tau > T_n(\nu_2)] &\leq \mathbb{P}_i^0[Y(T_n(\nu_2)) \leq C] \\
&= \mathbb{P}^0[Y_i(T_n(\nu_2)) - \mathbb{E}^0[Y(T_n(\nu_2))] \leq C - n\beta_i^{0,(n)}(\nu_2)] \\
&\leq \mathbb{P}^0[Y_i(T_n(\nu_2)) - \mathbb{E}^0[Y(T_n(\nu_2))] \leq n((C+1)/n - \beta^{(n)}(\nu_2)] \\
&\leq e^{-2n^{-1}(C+1-n\beta^{(n)}(\nu_2)} = e^{-2n((C+1)/n-\beta^{(n)}(\nu_2))^2} \qquad (24)
\end{aligned}$$

for $i = 1, \ldots, n$, where (24) holds from Hoeffding's inequality. Therefore, as $n \to \infty$,

$$\max_{1\leq i\leq n} \mathbb{P}_i^0[\tau > T_{\nu_2}^{(n)}] \leq \exp\left\{-2n\left[\frac{C+1}{n} - \beta^{(n)}(\nu_2)\right]^2\right\} \to 0.$$

□

We are now in position to prove Proposition 3.1.

Proof of Proposition 3.1. Let $\nu_1 < \nu_0 < \nu_2$. Observe from Lemma 3.9 that, as $n \to \infty$,

$$\mathbb{P}^0[\tau < T_n(\nu_1)] = \sum_{i=1}^n p_i^{(n)} \mathbb{P}_i^0[\tau < T_n(\nu_1)] \leq \max_{1\leq i\leq n} \mathbb{P}_i^0[\tau < T_n(\nu_1)] \to 0 \tag{25}$$

and

$$\mathbb{P}^0[\tau > T_n(\nu_2)] = \sum_{i=1}^n p_i^{(n)} \mathbb{P}_i^0[\tau > T_n(\nu_2)] \leq \max_{1\leq i\leq n} \mathbb{P}_i^0[\tau > T_n(\nu_2)] \to 0. \tag{26}$$

We have

$$\begin{aligned}
H^{\mathrm{LRU}} &= \mathbb{P}^0[Y_{X_0}(\tau) = 1] \\
&\geq \mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \geq T_n(\nu_1)] \\
&\geq \mathbb{P}^0[Y_{X_0}(T_n(\nu_1)) = 1, \tau \geq T_n(\nu_1)] \quad \text{from Lemma 3.8} \\
&\geq \mathbb{P}^0[Y_{X_0}(T_n(\nu_1) = 1] - \mathbb{P}^0[\tau < T_n(\nu_1)] \\
&= H^{TTL}(T_n(\nu_1) - \mathbb{P}^0[\tau < T_n(\nu_1)].
\end{aligned}$$

With the help of Lemmas 3.7 and (25) we find

$$\liminf_{n\to\infty} H^{\mathrm{LRU}} \geq \lim_{n\to\infty} H^{TTL}(T_n(\nu_1)) = h^\star(\nu_1).$$

Letting $\nu_1 \to \nu_0$,

$$\liminf_{n\to\infty} H^{\mathrm{LRU}} \geq h^\star(\nu_0).$$

For the other direction, note that

$$\begin{aligned}
H^{\mathrm{LRU}} &= \mathbb{P}^0[Y_{X_0}(\tau) = 1] \\
&\leq \mathbb{P}^0[Y_{X_0}(\tau) = 1, \tau \leq T_n(\nu_2)] + \mathbb{P}^0[\tau > T_n(\nu_2)] \\
&\leq \mathbb{P}^0[Y_{X_0}(T_n(\nu_2)) = 1, \tau \leq T_n(\nu_2)] + \mathbb{P}^0[\tau > T_n(\nu_2)] \quad \text{from Lemma 3.8} \\
&\leq \mathbb{P}^0[Y_{X_0}(T_n(\nu_2)) = 1] + \mathbb{P}^0[\tau > T_n(\nu_2)] \\
&= H^{TTL}(T_n(\nu_2)) + \mathbb{P}^0[\tau > T_n(\nu_2)].
\end{aligned}$$

With the help of Lemmas 3.7 and (26) we obtain

$$\limsup_{n\to\infty} H^{\mathrm{LRU}} \leq \lim_{n\to\infty} H^{TTL}(T_n(\nu_2)) = h^\star(\nu_2).$$

Letting $\nu_2 \to \nu_0$,

$$\limsup_{n\to\infty} H^{\mathrm{LRU}} \leq h^\star(\nu_0).$$

Therefore,

$$\lim_{n\to\infty} H^{\mathrm{LRU}} = h^\star(\nu_0).$$

□

4 Uniform convergence

It has been observed numerically in [10] that the characteristic time approximation is very accurate uniformly for contents of a wide range of popularity rank. In this section, we prove that this is indeed the case in the large system regime. Proposition 4.1 shows that the hit probabilities of individual contents in the characteristic time approximation converge uniformly to the corresponding hit probabilities in the LRU cache.

Recall that ν_0 is the unique root in $(0,\infty)$ of $h^\star(\nu) = \beta_0 \in (0,1)$ and that $T_n(\nu_0) = n\nu_0/\Lambda^{(n)}$.

Proposition 4.1. *Assume that $C \sim \beta_0 n$ with $\beta_0 \in (0,1)$. Suppose G is continuous. Then,*

$$\max_{1\leq i\leq n} \left|\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_n(\nu_0)) = 1]\right| \to 0 \quad \text{as } n \to \infty.$$

Proof. Let $\nu_1 < \nu_0 < \nu_2$. Note that

$$\begin{aligned}
\mathbb{P}_i^0[Y_i(\tau) = 1] &\geq \mathbb{P}_i^0[Y_i(\tau) = 1, \tau \geq T_n(\nu_1)] \\
&\geq \mathbb{P}_i^0[Y_i(T_n(\nu_1)) = 1, \tau \geq T_n(\nu_1)]] \quad \text{from Lemma 3.8} \\
&\geq \mathbb{P}_i^0[Y_i(T_n(\nu_1)) = 1] - \mathbb{P}_i^0[\tau < T_n(\nu_1)].
\end{aligned}$$

For the other direction, note that

$$\begin{aligned}
\mathbb{P}_i^0[Y_i(\tau) = 1] &\leq \mathbb{P}_i^0[Y_i(\tau) = 1, \tau \leq T_n(\nu_2)] + \mathbb{P}_i^0[\tau > T_n(\nu_2)] \\
&\leq \mathbb{P}_i^0[Y_i(T_n(\nu_2)) = 1, \tau \leq T_n(\nu_2)] + \mathbb{P}_i^0[\tau > T_n(\nu_2)] \quad \text{from Lemma 3.8} \\
&\leq \mathbb{P}_i^0[Y_i(T_n(\nu_2)) = 1] + \mathbb{P}_i^0[\tau > T_n(\nu_2)].
\end{aligned}$$

Since

$$\mathbb{P}_i^0[Y_i(T_n(\nu_1)) = 1] \leq \mathbb{P}_i^0[Y_i(T_n(\nu_0)) = 1] \leq \mathbb{P}_i^0[Y_i(T_n(\nu_2)) = 1],$$

we obtain

$$\begin{aligned}
&\left|\mathbb{P}_i^0[Y_i(\tau)=1]-\mathbb{P}_i^0[Y_i(T_n(\nu_0))=1]\right| \\
&\leq \mathbb{P}_i^0[Y_i(T_n(\nu_2))=1]-\mathbb{P}_i^0[Y_i(T_n(\nu_1))=1]+\mathbb{P}_i^0[\tau>T_n(\nu_2)]+\mathbb{P}_i^0[\tau<T_n(\nu_1)] \\
&= G_i(T_n(\nu_2))-G_i(T_n(\nu_1))+\mathbb{P}_i^0[\tau>T_n(\nu_2)]+\mathbb{P}_i^0[\tau<T_n(\nu_1)] \\
&= G(F'(\xi_i^{(n)})\nu_2)-G(F'(\xi_i^{(n)})\nu_1)+\mathbb{P}_i^0[\tau>T_n(\nu_2)]+\mathbb{P}_i^0[\tau<T_n(\nu_1)] \quad \text{by using (4).}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\max_{1\leq i\leq n}\left|\mathbb{P}_i^0[Y_i(\tau)=1]-\mathbb{P}_i^0[Y_i(T_n(\nu_0))=1]\right| \\
&\leq \max_{1\leq i\leq n}\left[G(F'(\xi_i^{(n)})\nu_2)-G(F'(\xi_i^{(n)})\nu_1)\right]+\max_{1\leq i\leq n}\mathbb{P}_i^0[\tau>T_n(\nu_2)]+\max_{1\leq i\leq n}\mathbb{P}_i^0[\tau<T_n(\nu_1)] \\
&\leq \sup_{y\geq 0}\left[G(y\nu_2)-G(y\nu_1)\right]+\max_{1\leq i\leq n}\mathbb{P}_i^0[\tau>T_n(\nu_2)]+\max_{1\leq i\leq n}\mathbb{P}_i^0[\tau<T_n(\nu_1)].
\end{aligned}$$

By Lemma 3.9,

$$\lim_{n\to\infty}\max_{1\leq i\leq n}\left|\mathbb{P}_i^0[Y_i(\tau)=1]-\mathbb{P}_i^0[Y_i(T_n(\nu_0)=1]\right|\leq\sup_{y\geq 0}\left[G(y\nu_2)-G(y\nu_1)\right].$$

The conclusion follows since the r.h.s. can be made arbitrarily small. Indeed, let ν_1, ν_2 be close enough to ν_0 so that $\nu_1, \nu_2 \in [\frac{1}{2}\nu_0, 2\nu_0]$. Given any $\epsilon > 0$, since $\nu_0 > 0$, there exists a large enough L such that

$$1-G(L\nu_0/2)<\epsilon.$$

Thus

$$\sup_{y\geq L}\left[G(y\nu_2)-G(y\nu_1)\right]\leq 1-G(L\nu_0/2)<\epsilon.$$

Being continuous, G is uniformly continuous on $[0, 2L\nu_0]$. When $\nu_2-\nu_1$ are small enough,

$$\sup_{y\in[0,L]}\left[G(y\nu_2)-G(y\nu_1)\right]<\epsilon.$$

Therefore,

$$\sup_{y\geq 0}\left[G(y\nu_2)-G(y\nu_1)\right]<\epsilon,$$

which concludes the proof. □

Proposition 4.1 provides an alternative proof of (11) in Proposition 3.1 since, by (9),

$$\begin{aligned}
\left|H^{\mathrm{LRU}}-H^{\mathrm{TTL}}(T_n(\nu_0))\right| &= \left|\sum_{i=1}^n p_i^{(n)}\left(\mathbb{P}_i^0[Y_i(\tau)=1]-\mathbb{P}_i^0[Y_i(T_n(\nu_0))=1]\right)\right| \\
&\leq \max_{1\leq i\leq n}\left|\mathbb{P}_i^0[Y_i(\tau)=1]-\mathbb{P}_i^0[Y_i(T_n(\nu_0))=1]\right|\to 0 \quad \text{as } n\to\infty.
\end{aligned}$$

Note, however, that (11) is proved in Section 3 without the additional assumption, used in Proposition 4.1, that G is continuous.

5 Rate of Convergence

Let $\delta_i = 0$ if the request process for content $i = 1,\ldots,n$ is Poisson and $\delta_i = 1$ otherwise. Let $\delta = \max_{1\leq i\leq n}\delta_i$. Throughout this section, we assume the aggregate request rate is normalized, i.e. $\sum_i \lambda_i = 1$, which can be achieved by changing the unit of time.

From $C(0) = 0$, $C(\infty) = n$ and the strict increasingness of the mapping $T \mapsto C(T)$ (see Lemma 3.3), we know that the equation $C(T) = C \in (0, n)$ has a unique solution in $(0,\infty)$, which we call T_0. Define $\mu_0 = C'(T_0) = \sum_{i=1}^n \lambda_i(1-G_i(T_0))$ (use (15)) the miss rate in a TTL cache with timer T_0. Assume that $\mu_0 > 0$.

5.1 Main Results

Proposition 5.1. *Suppose there exist a constant B and $\rho \in [\frac{\delta}{\mu_0 T_0}, 1]$ such that*

$$|\bar{G}_i(T_0) - \bar{G}_i(T_0 \pm \varepsilon T_0)| \leq B\varepsilon, \quad \textit{for } \varepsilon \in \left[\frac{\delta}{\mu_0 T_0}, \rho\right]. \tag{27}$$

Let $D_0 = \sqrt{\frac{2}{n}}\mu_0 T_0$. Assume $D_0/B \geq \sqrt{\frac{e}{2}}$ and

$$1 + \sqrt{\log \frac{D_0}{B}} \leq D_0 \rho - \sqrt{\frac{2}{n}}\delta. \tag{28}$$

Then,

$$\max_{1 \leq i \leq n} \left|\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1]\right| \leq \frac{B\delta}{\mu_0 T_0} + \frac{B}{D_0}\left(\sqrt{\log \frac{D_0}{B}} + 1 + \frac{1}{\sqrt{2}}\right). \tag{29}$$

In particular,

$$|H^{\mathrm{LRU}} - H^{TTL}(T_0)| \leq \frac{B\delta}{\mu_0 T_0} + \frac{B}{D_0}\left(\sqrt{\log \frac{D_0}{B}} + 1 + \frac{1}{\sqrt{2}}\right). \tag{30}$$

Note that (27) holds for a large class of distributions.

Example 5.2. For Poisson arrivals, $\bar{G}_i(t) = e^{-\lambda_i t}$. For any $\varepsilon \geq 0$,

$$0 \leq \bar{G}_i(T_0) - \bar{G}_i(T_0 + \varepsilon T_0) = e^{-\lambda_i T_0}(1 - e^{-\lambda_i \varepsilon T_0}) \leq \lambda_i T_0 e^{-\lambda_i T_0}\varepsilon \leq e^{-1}\varepsilon.$$

For $\varepsilon \in [0, 1]$,

$$0 \leq \bar{G}_i(T_0 - \varepsilon T_0) - \bar{G}_i(T_0) \leq \sup_{x \geq 0} e^{-x}(e^{\varepsilon x} - 1) = (1 - \varepsilon)^{\frac{1}{\varepsilon} - 1}\varepsilon \leq \varepsilon.$$

Thus (27) holds with $B = 1$ and $\rho = 1$.

Example 5.3. Suppose G_i's are continuously differentiable. By the mean value theorem, there exists $\xi_i \in [1, 1 + \varepsilon]$ such that

$$0 \leq \bar{G}_i(T_0) - \bar{G}_i(T_0 + \varepsilon T_0) = G_i'(\xi_i T_0)\varepsilon T_0 \leq \xi_i T_0 G_i'(\xi_i T_0)\varepsilon \leq \left[\sup_{t \geq 0} t G_i'(t)\right]\varepsilon,$$

Similarly, there exists $\zeta_i \in [1 - \varepsilon, 1]$ such that

$$0 \leq \bar{G}_i(T_0 - \varepsilon T_0) - \bar{G}_i(T_0) = G_i'(\zeta_i T_0)\varepsilon T_0 \leq \frac{\zeta_i}{1 - \varepsilon} T_0 G_i'(\zeta_i T_0)\varepsilon \leq \frac{\varepsilon}{1 - \rho}\left[\sup_{t \geq 0} t G_i'(t)\right].$$

Thus (27) holds with $\rho \in [\frac{\delta}{\mu_0 T_0}, 1)$ and

$$B = \frac{1}{1 - \rho} \max_{1 \leq i \leq n} \sup_{t \geq 0} t G_i'(t).$$

Note that $\sup_{t \geq 0} t G_i'(t) < \infty$ is invariant under arbitrary rescaling, so we may replace G_i by its scaled version for convenience. Since G_i has finite mean, $\sup_{t \geq 0} t G_i'(t) < \infty$, and hence $B < \infty$. Note that B may diverge to infinity as n increases. If the G_i's are from the same scale family, i.e. $G_i(t) = G(\lambda_i t)$ for all i, however,

$$B = \frac{1}{1 - \rho} \max_{1 \leq i \leq n} \sup_{t \geq 0} \lambda_i t G'(\lambda_i t) = \frac{1}{1 - \rho} \sup_{t \geq 0} t G'(t) < \infty$$

for any n. In particular, for Poisson arrivals, (27) holds with $B = e^{-1}(1 - \rho)^{-1}$ and $\rho \in [0, 1)$.

An example where the G_i's are not from the same scale family but we still have finite B for all n is provide by gamma distributions with shape parameters α_i that have a common upper bound $\alpha_{\max}$, i.e.

$$G_i'(t) = \frac{1}{\Gamma(\alpha_i)} t^{\alpha_i - 1} e^{-t},$$

where we have set the scale parameter to 1 by the invariance mentioned in the previous paragraph. In this case,

$$\sup_{t \geq 0} t G_i'(t) = \frac{1}{\Gamma(\alpha_i)} \sup_{t \geq 0} t^{\alpha_i} e^{-t} = \frac{\alpha_i^{\alpha_i} e^{-\alpha_i}}{\Gamma(\alpha_i)},$$

and hence

$$B = \frac{1}{1-\rho} \max_{1 \leq i \leq n} \frac{\alpha_i^{\alpha_i} e^{-\alpha_i}}{\Gamma(\alpha_i)} \leq \frac{1}{1-\rho} \max_{0 < \alpha \leq \alpha_{\max}} \frac{\alpha^{\alpha} e^{-\alpha}}{\Gamma(\alpha)} < \infty$$

for any n.

Corollary 5.4. *Assume that $\delta_i = 0$ for all $i = 1, \ldots$ (Poisson requests). If $C \sim \beta_0 n$ with $\beta_0 \in (0,1)$ and $\min_i \lambda_i = \Omega(n^{-\gamma})$ for $1 \leq \gamma < 3/2$, then*

$$\max_{1 \leq i \leq n} \left| \mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1] \right| = O\left(n^{\gamma - 3/2} \sqrt{\log n} \right).$$

Note that for Zipf's distribution with parameter α, i.e. $\lambda_i \propto i^{-\alpha}$,

$$\min_i \lambda_i = \begin{cases} \Theta(n^{-1}), & 0 \leq \alpha < 1; \\ \Theta(n^{-1}/\log n), & \alpha = 1; \\ \Theta(n^{-\alpha}), & \alpha > 1. \end{cases}$$

Thus for $\alpha < 3/2$, the condition is satisfied with $\gamma = \max\{1, \alpha\}$.

Proof. From Example 5.2 we know that (27) holds with $B = \rho = 1$, which by Proposition 5.1 gives

$$\max_{1 \leq i \leq n} \left| \mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1] \right| \leq \frac{\sqrt{n}}{\sqrt{2}\mu_0 T_0} \left(\sqrt{\log\left(\sqrt{\frac{2}{n}} \mu_0 T_0 \right) + 1} + \frac{1}{\sqrt{2}} \right). \tag{31}$$

Note that (Hint: $G_i = \hat{G}_i$ when $\delta_i = 0$)

$$\mu_0 = \sum_{i=1}^{n} \lambda_i \bar{G}_i(T_0) \geq \min_i \lambda_i \sum_{i=1}^{n} \bar{G}_i(T_0) = (n - C) \min_i \lambda_i = \Omega(n^{1-\gamma}).$$

On the other hand $C = \sum_{i=1}^n \hat{G}_i(T_0) \leq T_0 \Lambda^{(n)}$, so that $T_0 = \Omega(n)$, $\mu_0 T_0 / \sqrt{n} = \Omega(n^{3/2-\gamma})$ and therefore $\sqrt{n}/\mu_0 T_0 = O(n^{\gamma - 3/2})$, which concludes the proof.

□

5.2 Proof of Proposition 5.1

Define $C_i^0(T) = \mathbb{E}_i^0[Y(T)]$, the expected number of distinct requests in $[-T, 0)$ given that a request for content i is made at $t = 0$.

Lemma 5.5. *For $T > 0$*

$$|C_i^0(T) - C(T)| \leq \delta_i, \tag{32}$$

for $i = 1, \ldots, n$.

Proof.

$$\begin{aligned} C_i^0(T) &= \mathbb{E}_i^0[Y(T)] = \sum_{j=1}^n \mathbb{P}_i^0(Y_j(t)=1) = \sum_{j=1, j\neq i}^n \hat{G}_j(T) + G_i(T) \quad \text{from Lemma 3.2} \\ &= C(T) + G_i(T) - \hat{G}_i(T), \end{aligned}$$

by using Lemma 3.3. If $\delta_i = 0$ then $G_i(T) = \hat{G}_i(T)$ and $C_i^0(T) = C(T)$ which shows (32). Assume that $\delta_i \neq 0$. We have

$$|C_i^0(T) - C(T)| = |G_i(T) - \hat{G}_i(T)| \leq 1$$

since G_i and $\hat{G}_i$ are cdfs. This proves the lemma. □

Lemma 5.6. *For $\varepsilon \geq \delta_i/(\mu_0 T_0)$, we have*

$$\mathbb{P}_i^0[\tau > (1+\varepsilon)T_0] \leq e^{-2n^{-1}(\mu_0 T_0 \varepsilon - \delta_i)^2},$$

and

$$\mathbb{P}_i^0[\tau < (1-\varepsilon)T_0] \leq e^{-2n^{-1}(\mu_0 T_0 \varepsilon - \delta_i)^2}.$$

Proof. Recall the definition of τ,

$$\tau = \inf\{t : Y(t) \geq C\}.$$

Let $T_1 = (1-\varepsilon)T_0$. By Lemma 5.5

$$C - C_i^0(T_1) \geq C - \delta_i - C(T_1) = C(T_0) - C(T_1) - \delta_i.$$

By the concavity of $C(T)$ (see Lemma 3.3)

$$C(T_0) - C(T_1) \geq C'(T_0)(T_0 - T_1) = \mu_0 T_0 \varepsilon \geq \delta_i$$

so that

$$C(T_0) - C(T_1) - \delta_i > \mu_0 T_0 \varepsilon - \delta_i \geq 0.$$

By Hoeffding's inequality,

$$\begin{aligned} \mathbb{P}_i^0[\tau < T_1] &\leq \mathbb{P}_i^0[Y(T_1) \geq C] \\ &= \mathbb{P}_i[Y(T_1) - \mathbb{E}_i^0[Y(T_1)] \geq C - C_i^0(T_1)] \quad \text{as } \mathbb{E}_i^0[Y(T_1)] = C_i^0(T_1) \\ &\leq \mathbb{P}_i^0[Y(T_1) - \mathbb{E}_i^0[Y(T_1)] \geq \mu_0 T_0 \varepsilon - \delta_i] \\ &\leq e^{-2n^{-1}(\mu_0 T_0 \varepsilon - \delta_i)^2} \quad \text{by Hoeffding's inequality.} \end{aligned}$$

Similarly, for $T_2 = (1+\varepsilon)T_0$,

$$C - C_i^0(T_2) \leq C - C(T_2) + \delta_i \leq \delta_i - C'(T_0)(T_2 - T_0) = \delta_i - \mu_0 T_0 \varepsilon \leq 0,$$

so that

$$\begin{aligned} \mathbb{P}_i^0[\tau > T_2] &\leq \mathbb{P}_i^0[Y(T_2) < C] \\ &= \mathbb{P}_i^0[Y(T_2) - \mathbb{E}_i^0[Y(T_1)] < C - C_i^0(T_2)] \quad \text{as } \mathbb{E}_i^0[Y(T_2)] = C_i^0(T_2) \\ &\leq \mathbb{P}_i^0[Y(T_2) - \mathbb{E}_i^0[Y(T_1)] < \delta_i - \mu_0 T_0 \varepsilon] \\ &\leq e^{-2n^{-1}(\mu_0 T_0 \varepsilon - \delta_i)^2} \text{ by Hoeffding's inequality.} \end{aligned}$$

□

We are now in position to prove Proposition 5.1.

Proof of Proposition 5.1. Fix $i \in \{1, \ldots, n\}$. Let $T_1 = (1-\varepsilon)T_0$ and $T_2 = (1+\varepsilon)T_0$, where $\varepsilon \geq \varepsilon_i := \delta_i/(\mu_0 T_0)$. Let

$$U(i, \varepsilon) = B\varepsilon + e^{-2n^{-1}(\mu_0 T_0 \varepsilon - \delta_i)^2}.$$

Note that

$$\begin{aligned}\mathbb{P}_i^0[Y_i(\tau) = 1] &\geq \mathbb{P}_i^0[Y_i(\tau) = 1, \tau \geq T_1] \\ &\geq \mathbb{P}_i^0[Y_i(T_1) = 1, \tau \geq T_1] \quad \text{from Lemma 3.8} \\ &\geq \mathbb{P}_i^0[Y_i(T_1) = 1] - \mathbb{P}_i^0[\tau < T_1].\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}_i^0[Y_i(T_0) = 1] - \mathbb{P}_i^0[Y_i(\tau) = 1] &\leq \bar{G}_i(T_0 - \varepsilon T_0) - \bar{G}_i(T_0) + e^{-2n^{-1}(\mu_0 T_0 \varepsilon - \delta_i)^2} \\ &\leq U(i, \varepsilon).\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}_i^0[Y_i(\tau) = 1] &\leq \mathbb{P}_i^0[Y_i(\tau) = 1, \tau \leq T_2] + \mathbb{P}_i^0[\tau > T_2] \\ &\leq \mathbb{P}_i^0[Y_i(T_2) = 1, \tau \leq T_2] + \mathbb{P}_i^0[\tau > T_2] \quad \text{from Lemma 3.8} \\ &\leq \mathbb{P}_i^0[Y_i(T_2) = 1] + \mathbb{P}_i^0[\tau > T_2].\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1] &\leq \bar{G}_i(T_0) - \bar{G}_i(T_0 + \varepsilon T_0) + e^{-2n^{-1}(\mu_0 T_0 \varepsilon - \delta_i)^2} \\ &\leq U(i, \varepsilon).\end{aligned}$$

Therefore,

$$\left|\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1]\right| \leq U(i, \varepsilon).$$

Since the above inequality holds for any $\varepsilon \in [\varepsilon_i, \rho]$,

$$\left|\mathbb{P}_i^0[Y_i(\tau) = 1] - \mathbb{P}_i^0[Y_i(T_0) = 1]\right| \leq \inf_{\varepsilon \in [\varepsilon_i, \rho]} U(i, \varepsilon).$$

Let $\eta = \sqrt{\frac{2}{n}}(\mu_0 T_0 \varepsilon - \delta_i)$, $D_0 = \sqrt{\frac{2}{n}}\mu_0 T_0$, and $A_0 = D_0 \rho - \sqrt{\frac{2}{n}}\delta$. In this notation, $U(i, \varepsilon) = B\varepsilon + e^{-\eta^2} = \frac{B\delta_i}{\mu_0 T_0} + \frac{B}{D_0}\eta + e^{-\eta^2}$. For $\varepsilon \in [\varepsilon_i, \rho]$, $\eta \in [0, D_0\rho - \sqrt{\frac{2}{n}}\delta_i] \supset [0, A_0]$. Thus

$$\inf_{\varepsilon \in [\varepsilon_i, \rho]} U(i, \varepsilon) = \frac{B\delta_i}{\mu_0 T_0} + \inf_{\eta \in [0, D_0\rho - \sqrt{\frac{2}{n}}\delta_i]} \left\{\frac{B}{D_0}\eta + e^{-\eta^2}\right\} \leq \frac{B\delta_i}{\mu_0 T_0} + \inf_{\eta \in [0, A_0]} \left\{\frac{B}{D_0}\eta + e^{-\eta^2}\right\}.$$

The stationary point of the function $\eta \mapsto B\eta/D_0 + e^{-\eta^2}$ satisfies

$$2\eta e^{-\eta^2} = B/D_0. \tag{33}$$

Note that the function $\eta \mapsto 2\eta e^{-\eta^2}$ maps $[0, \infty)$ onto $[0, \sqrt{2/e}]$. Since $D_0/B \geq \sqrt{\frac{e}{2}}$ by assumption, the above equation has two positive roots (they are identical if $D_0/B = \sqrt{e/2}$). Let η_0 be the larger root ($\geq 1/\sqrt{2}$), which minimizes $B\eta/D_0 + e^{-\eta^2}$. Note that

$$B/D_0 = 2\eta_0 e^{-\eta_0^2} \leq e^{2\eta_0 - 1} e^{-\eta_0^2} = e^{-(\eta_0 - 1)^2},$$

which implies

$$\eta_0 \leq 1 + \sqrt{\log \frac{D_0}{B}}.$$

By (28), this implies $\eta_0 \in [0, A_0]$. Thus,

$$\inf_{\eta \in [0,A_0]} \left\{ \frac{B}{D_0}\eta + e^{-\eta^2} \right\} = \frac{B}{D_0}\eta_0 + e^{-\eta_0^2} = \frac{B}{D_0}\left(\eta_0 + \frac{1}{2\eta_0}\right) \leq \frac{B}{D_0}\left(\sqrt{\log \frac{D_0}{B}} + 1 + \frac{1}{\sqrt{2}}\right),$$

where the second equality follows from (33), and the last inequality holds because $\eta \mapsto \eta + 1/(2\eta)$ is increasing for $\eta \geq 1/\sqrt{2}$ and $\eta_0 \geq 1/\sqrt{2}$. This shows (29).

The last statement (30) follows from (29) and the inequality

$$|H^{LRU} - H^{TTL}(T_0)| = |\sum_{i=1}^{n} p_i^{(n)} (\mathbb{P}_i^0[Y(\tau)] - P_i^0[Y(T_0)])| \leq |\mathbb{P}_i^0[Y(\tau)] - P_i^0[Y(T_0)]|.$$

□

6 Extension to several content popularity cdfs

In this section, we consider the situation where $K \geq 1$ service providers (SP) share a common LRU cache.

We assume that there are nb_k documents associated with SP $k = 1, \ldots, K$. Successive requests to content i associated with SP k follow a stationary and ergodic process with cdf $G_{k,i}$ and intensity $\lambda_{k,i} > 0$, for $i = 1, \ldots, nb_k$. All these $n(b_1 + \cdots + b_K)$ stationary and ergodic processes are assumed to be mutually independent.

Notation and assumption below hold for each $k = 1, \ldots, K$. Let $\Lambda_k^{(n)} = \sum_{i=1}^{nb_k} \lambda_{k,i}$ be the total request rate of contents associated with SP k. We assume that there exists a continuously differentiable cdf F_k such that

$$p_{k,i}^{(n)} := \frac{\lambda_{k,i}}{\Lambda_k^{(n)}} = F_k\left(\frac{i}{nb_k}\right) - F_k\left(\frac{i-1}{nb_k}\right) = \frac{1}{nb_k}F_k'(\xi_{k,i}^{(n)}), \quad i = 1, \ldots, nb_k,$$

where $\xi_{k,i}^{(n)} \in \left(\frac{i-1}{nb_k}, \frac{i}{nb_k}\right)$. We assume that $F_k'(x) > 0$ a.e. on $[0, 1]$ and allow $F_k'(0)$ to be infinite.

We assume that

$$G_{k,i}(x) = G_k(\lambda_{k,i}x),$$

for some CDF G_k with mean 1. Last, we assume that $\lim_{n\to\infty} \Lambda_k^{(n)} = \Lambda_k \in (0, \infty)$.

The following result holds:

Proposition 6.1. *Assume that $C \sim n\beta_0$ with $\beta_0 \in (0, 1)$, and define $T_n(\nu) = \nu n / \sum_{k=1}^{K} \Lambda_k^{(n)}$. Then,*

$$\lim_{n\to\infty} H^{\mathrm{LRU}} = \lim_{n\to\infty} H^{\mathrm{TTL}}(T_n(\nu_0)) \tag{34}$$

$$= \sum_{k=1}^{K} a_k \int_0^1 F_k'(x) G_k(\nu_0 a_k b_k^{-1} F_k'(x)) dx, \tag{35}$$

where ν_0 is the unique solution in $(0, \infty)$ of the equation

$$\sum_{k=1}^{K} \frac{b_k}{B_K} \int_0^1 \hat{G}_k(\nu a_k b_k^{-1} F_k'(x)) dx = \beta_0,$$

with $a_k := \Lambda_k / \sum_{j=1}^{K} \Lambda_j$ and $B_K := \sum_{k=1}^{K} b_k$.

Proof. Consider a TTL cache with timer $T_n(\nu) = \nu n / \sum_{k=1}^{K} \Lambda_k^{(n)}$ where $\nu > 0$ is arbitrary. This cache receives requests for nB_K contents.

Define $B_k := \sum_{j=1}^{k} b_k$ for $k = 1, \ldots, K$, with $B_0 = 1$ by convention. Label the contents from 1 to nB_K, where contents $nB_{k-1} + 1, \ldots, nB_k$ belong to SP k.

Lemmas 3.2-3.3, 3.5-3.6 and 3.8-3.9 apply to independent and stationary request processes $1, \ldots, nB_K$ by replacing n by nB_k, G_i by $G_{k,i-nB_{k-1}}$ and $p_i^{(n)}$ by $\lambda_{k,i}/\sum_{k=1}^{K}\Lambda_k^{(n)}$ if $nB_{k-1} + 1 \le i \le nB_k$. Let us now focus on Lemmas 3.4 and 3.7.

Define $a_k^{(n)} = \Lambda_k^{(n)}/\sum_{j=1}^{K}\Lambda_j^{(n)}$ and $a_k = \lim_{n\to\infty} a_k^{(n)} = \Lambda_k/\sum_{j=1}^{K}\Lambda_j$. Mimicking the first steps of the proof of Lemma 3.7 gives

$$\begin{aligned}
H^{0,(nB_K)}(\nu) = \mathbb{P}^0[Y_{X_0}(T_n(\nu)) = 1] &= \sum_{i=1}^{nB_K} \mathbb{P}^0[X_0 = i]\mathbb{P}^0[Y_i(T_n(\nu)) = 1 \mid X_0 = i] \\
&= \sum_{k=1}^{K}\sum_{i=1}^{nb_k} \frac{\lambda_{k,i}^{(n)}}{\sum_{k=1}^{K}\Lambda_k^{(n)}} G_{k,i}(T_n(\nu)) = \sum_{k=1}^{K} a_k^{(n)} \sum_{i=1}^{nb_k} p_{k,i}^{(n)} G_k(\lambda_{k,i}T_n(\nu)) \\
&= \sum_{k=1}^{K} a_k^{(n)} \sum_{i=1}^{nb_k} \frac{1}{nb_k} F_k'(\xi_{k,i}^{(n)}) G_k\left(\nu a_k^{(n)} b_k^{-1} F_k'(\xi_{k,i}^{(n)})\right) \\
&\to \sum_{k=1}^{K} a_k \int_0^1 F_k'(x) G_k(\nu a_k b_k^{-1} F_k'(x))dx \quad \text{as } n \to \infty,
\end{aligned} \tag{36}$$

since for each $k = 1, \ldots, K$ the cdf F_k is continuously differentiable in $(0, 1)$.

Similarly (see Lemma 3.4)

$$\begin{aligned}
\beta^{(nB_K)}(\nu) &= \frac{1}{nB_K} \sum_{k=1}^{K}\sum_{i=1}^{nb_k} \hat{G}_{k,i}(T_n(\nu)) \\
&= \sum_{k=1}^{K} \frac{b_k}{B_K} \frac{1}{nb_k} \sum_{i=1}^{nb_k} \hat{G}_k(\lambda_{k,i}T_n(\nu)) \\
&= \sum_{k=1}^{K} \frac{b_k}{B_K} \frac{1}{nb_k} \sum_{i=1}^{nb_k} \hat{G}_k\left(\nu a_k^{(n)} b_k^{-1} F_k'(\xi_{k,i}^{(n)})\right) \\
&\to \sum_{k=1}^{K} \frac{b_k}{B_K} \int_0^1 \hat{G}_k(\nu a_k b_k^{-1} F_k'(x))dx \quad \text{as } n \to \infty.
\end{aligned} \tag{37}$$

The second part of Lemma 3.4 also holds (cf. (19)) since $F_k'(x) > 0$ a.e. on $[0, 1]$ for all $k = 1, \ldots, K$.

The rest of the proof mimics that of Proposition 3.1 with $h^\star(\nu)$ and $\beta^\star(\nu)$ given by the r.h.s of (36) and (37), respectively, and is omitted. □

7 Conclusions

In this paper, we developed an approximation for the aggregate and individual content hit rates of an LRU cache fir the case that content requests are described by independent stationary and ergodic processes. This approximation extends one first proposed and studied by Fagin [7] for the independent reference model and provides the theoretical basis for approximations introduced in [11] for content requests described by independent renewal processes. We showed that the approximations become exact in the limit as the number of contents goes to infinity while the ratio of this and the cache size remains constant. Last, we established the rate of convergence for the approximation as number of contents increases.

Future directions include extension of these results to other cache policies such as FIFO and random and to networks of caches. In addition, it is desirable to relax independence between different content request streams.

References

[1] B. Baccelli and P. Brémaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, volume 26 of *Applications of Mathematics, Stochatic Modelling and Applied Probability.* Springer-Verlag Berlin Heidelberg, 2nd edition, 2003.

[2] J. R. Bitner. Heuristics that monotonically organize data structures. *SIAM J. Computing*, 8:82–110, 1979.

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

[4] H. Che, Y. Tung, and Z. Wang. Hierarchical web caching systems: Modeling, design and experimental results. *IEEE Journal on Selected Areas in Communications*, 20(7):1305–1314, 2002.

[5] N. Choungmo Fofack, P. Nain, G. Neglia, and D. Towsley. Performance evaluation of hierarchical ttl-based cache networks. *Computer Networks*, 65:212–231, June 2014.

[6] E. G. Coffman and P. Jelenkovic. Performance of the move-to-front algorithm with markov-modulated request sequences. *Operations Research Letters*, 25:109–118, 1999.

[7] R. Fagin. Asymptotic miss ratios over independent references. *Journal of Computer and System Sciences*, 14(2):222–250, 1977.

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

[9] G. B. Folland. *Real analysis: modern techniques and their applications.* John Wiley & Sons, 2013.

[10] C. Fricker, P. Robert, and J. Roberts. A versatile and accurate approximation for lru cache performance. In *Proceedings of the 24th International Teletraffic Congress (ITC 24)*, Kraków, Poland, September 4-7 2012.

[11] M. Garetto, E. Leonardi, and V. Martina. A unified approach to the performance analysis of caching systems. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, 1(3), May 2016.

[12] N. Gast and B. Van Houdt. Asymptotically exact ttl-approximations of the cache replacement algorithms lru (m) and h-lru. In *Proceedings of the 28th International Teletraffic Congress (ITC 28)*, volume 1, pages 157–165, Würzburg, Germany, September 12-16 2016. IEEE.

[13] P. Jelenkovic and A. Radovanović. Least-recently used caching with dependent requests. *Theoretical Computer Science*, 326:293–327, 2004.

[14] P. Jelenkovic, A. Radovanović, and M. Squillante. Critical sizing of lru caches with dependent requests. *J. of Applied Probability*, 43(4):1013–1027, 2006.

[15] W. F. King. Analysis of demand paging algorithm. *Information Processing*, 71:485–490, 1972.

[16] E. Leonardi and G. L. Torrisi. Modeling least recently used caches with shot noise request processes. *SIAM Journal on Applied Mathematics*, 77(2):361–383, 2017.

[17] E. J. Rosensweig, J. Kurose, and D. Towsley. Approximate models for general cache networks. In *Proceedings of Infocom 2010*, pages 1100–1108, San Diego, CA, USA, March 14-19 2010.