HAL
open science

On the Convergence of the TTL Approximation for an LRU Cache under Independent Stationary Request Processes

Bo Jiang, Philippe Nain, Don Towsley

To cite this version:

Bo Jiang, Philippe Nain, Don Towsley. On the Convergence of the TTL Approximation for an LRU Cache under Independent Stationary Request Processes. ACM Transactions on Modeling and Performance Evaluation of Computing Systems, 2018, 3 (4), 10.1145/3239164 . hal-01673272v4

HAL Id: hal-01673272

https://inria.hal.science/hal-01673272v4

Submitted on 10 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the Convergence of the TTL Approximation for an LRU Cache under Independent Stationary Request Processes

BO JIANG, University of Massachusetts Amherst
PHILIPPE NAIN, Inria
DON TOWSLEY, University of Massachusetts Amherst

The modeling and analysis of an LRU cache is extremely challenging as exact results for the main performance metrics (e.g. hit rate) are either lacking or cannot be used because of their high computational complexity for large caches. As a result, various approximations have been proposed. The state-of-the-art method is the so-called TTL approximation, first proposed and shown to be asymptotically exact for IRM requests by Fagin [14]. It has been applied to various other workload models and numerically demonstrated to be accurate but without theoretical justification. In this paper we provide theoretical justification for the approximation in the case where distinct contents are described by independent stationary and ergodic processes. We show that this approximation is exact as the cache size and the number of contents go to infinity. This extends earlier results for the independent reference model. Moreover, we establish results not only for the aggregate cache hit probability but also for every individual content. Last, we obtain bounds on the rate of convergence.

CCS Concepts: •**Mathematics of computing → Stochastic processes;** •**Networks → Network performance modeling;** •**Theory of computation → Caching and paging algorithms;**

Additional Key Words and Phrases: Cache, LRU, Characteristic time, TTL approximation, Stationary request processes, Convergence, Asymptotic exactness

1 INTRODUCTION

Caches are key components 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 have been introduced and analyzed over the last few decades, mostly based on the least recently used algorithm (LRU). Considerable work has focused on analyzing these policies [6, 7, 11, 16, 21, 22, 24]. Since exact results for the main performance metrics (e.g. hit rate) are either lacking or cannot be used because of their high computational complexity for large caches, approximations have been proposed [8, 12, 14, 20, 23, 27, 28]. Of all the approximation techniques developed, the state of the art is provided by the so-called TTL approximation based on time-to-live (TTL) caches, which has been demonstrated to be accurate for various caching policies and traffic models [5, 8, 10, 13, 14, 17–19, 25]. In this paper, we focus on the TTL approximation for the LRU cache with stationary requests. In a TTL cache, a time-to-live timer is set to its maximum value T each time the content is requested. The content is evicted from the cache when the timer expires.

This research was sponsored by the U.S. ARL and the U.K. MoD under Agreement Number W911NF-16-3-0001 and by the NSF under Grants CNS-1413998 and CNS-1617437. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

The link between an LRU cache and a TTL[1] cache was first pointed out in [14] for i.i.d. requests (the so-called independence reference model - IRM). 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 [16], this work went unnoticed and [8] reintroduced the approximation, without theoretical justification, for LRU under Poisson requests. Fricker et al [17] provided some theoretical justification for the approximation by establishing a central limit theorem of the characteristic time under Poisson requests (see Remark 3 in Section 4.2 for a brief discussion). More recently, [18] extended the TTL 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. For independent Markovian Arrival Processes, [19] developed TTL approximations for the more complicated LRU(m) and h-LRU policies, both including LRU as a special case. All the aforementioned work focused on stationary request processes with no dependence between different contents. Dependent and so-called time-asymptotically stationary requests were considered in [27], but the results therein do not apply to the TTL approximation (see Section 4 for a brief discussion of this work). Non-stationary request processes were considered in [25], where a TTL approximation is developed for the hit probability in a single LRU cache and in a tandem of LRU caches, under the so-called shot noise request model. It is also shown in [25] that the cache eviction time converges to the characteristic time of the TTL approximation as the cache size goes to infinity.

The objective of the present paper is to provide a rigorous theoretical justification of the TTL approximation for LRU in [18] and its generalization to independent stationary content request processes. To the best of our knowledge, such a justification was only provided in [14], and later on in [20], under IRM (see Section 2.3 for a discussion of Theorem 1 in [20]).

We make the following contributions in this paper. First, we prove under the assumption that requests to distinct contents are described by mutually independent stationary and ergodic point processes, that the hit probability for each content under LRU converges to that for a TTL cache operating with a single timer value, called the LRU characteristic time, independent of the content. Moreover, we derive rates of convergence for individual content hit probabilities under LRU to those under TTL using the LRU characteristic time. Under additional mild conditions, we then derive expressions for the characteristic time and the aggregate hit probability in the limit as the cache size and the number of contents go to infinity. This last result extends the results of Fagin [14] for the independence reference model to a more general setting of independent stationary and ergodic content request processes.

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 main results of our paper. Section 4 proves the main result of the paper, namely the convergence of hit probabilities under LRU to those under TTL with bounds on the rate of convergence given in Section 5. Section 6 extends Fagin's results to the more general case of stationary and ergodic request processes. Last concluding statements are provided in Section 7.

[1]Fagin worked with the so-called working-set policy, which is the discrete time version of the TTL policy. The result can be easily translated into one for the TTL approximation - also referred to as the Che's approximation in the literature, following the work of Che et al. in [8] - under Poisson requests.

2 MODEL AND BACKGROUND

We introduce the model for content request processes in Section 2.1 and the content popularity in Section 2.2. Section 2.3 presents the TTL approximation that approximates hit probabilities of an LRU cache by those of a TTL cache with an appropriately chosen timer value.

2.1 Content Request Process

We consider a cache of size C_n serving n unit sized contents labelled $i = 1, \dots, n$, where $C_n \in (0, n)$. We assume that $C_n \to \infty$ as $n \to \infty$. In particular, several results will be obtained under the assumption that $C_n \sim \beta_0 n$ with $\beta_0 \in (0, 1)$. Requests for the contents are described by n independent stationary and ergodic simple point processes $N_{n,i} := \{t_{n,i}(k), k \in \mathbb{Z}\}$, where $-\infty \leq \cdots < t_{n,i}(-1) < t_{n,i}(0) \leq 0 < t_{n,i}(1) < \cdots \leq \infty$ represent successive request times to content $i = 1, \dots, n$. We assume the point processes are defined on a common probability space with probability measure $\mathbb{P}$ and associated expectation operator $\mathbb{E}$. Let $0 < \lambda_{n,i} < \infty$ denote the intensity of request process $N_{n,i}$, i.e., the long term average request rate for content i (see e.g. [3, Sections 1.1 and 1.6] for an introduction to stationary and ergodic point processes). Note that $\mathbb{P}[t_{n,i}(0) = 0] = 0$ for all i [3, Section 1.1.4], i.e. no request arrives precisely at time 0. The same request processes were considered in [15] for TTL caches.

Following [10], we will use Palm calculus for stationary and ergodic point processes [3]. Let $\mathbb{P}^0_{n,i}$ be the Palm probability[2] associated with the point process $N_{n,i}$ (see e.g. [3, Eq. (1.2.1)]). In particular, $\mathbb{P}^0_{n,i}[t_{n,i}(0) = 0] = 1$, i.e. under $\mathbb{P}^0_{n,i}$ content i is requested at time $t = 0$. It is known that [3, Exercice 1.2.1]

$$\mathbb{E}^0_{n,i}[t_{n,i}(1)] = \frac{1}{\lambda_{n,i}}, \tag{1}$$

where $\mathbb{E}^0_{n,i}$ is the expectation operator associated with $\mathbb{P}^0_{n,i}$. Define

$$G_{n,i}(t) = \mathbb{P}^0_{n,i}[t_{n,i}(1) \leq t], \tag{2}$$

the cdf of the inter-request time for content i under $\mathbb{P}^0_{n,i}$.

For any distribution F, we denote its mean by m_F and the corresponding ccdf by $\bar{F} := 1 - F$. For any F with support in $[0, \infty)$ and $m_F \in (0, \infty)$, we define an associated distribution $\hat{F}$ by

$$\hat{F}(t) = \frac{1}{m_F} \int_0^t \bar{F}(z)dz, \quad t \geq 0. \tag{3}$$

It is well-known that (see e.g. [3, Section 1.3.4])

$$\mathbb{P}[-t_{n,i}(0) \leq t] = \hat{G}_{n,i}(t) = \lambda_{n,i} \int_0^t \bar{G}_{n,i}(z)dz, \tag{4}$$

with $m_{G_{n,i}} = 1/\lambda_{n,i}$ from (1). Note that $\mathbb{P}[-t_{n,i}(0) \leq t]$ is the cdf of the time elapsed since content i was last requested before the random observation time $t = 0$ (recall that the system is in steady state at time $t = 0$), often referred to as the age distribution of the last request for content i.

We assume all cdfs $G_{n,i}$ are continuous. Let

$$G^*_{n,i}(t) = G_{n,i}(t/\lambda_{n,i}) \tag{5}$$

[2]Readers unfamiliar with Palm probability can think of $\mathbb{P}^0_{n,i}$ as being defined by $\mathbb{P}^0_{n,i}[A] = \mathbb{P}[A \mid t_{n,i}(0) = 0]$ for any event A, i.e. the conditional probability conditioned on the event that content i is requested at time 0, although the definition is more general.

be the scaled version of $G_{n,i}$ that is standardized in the sense that it has unit mean. We assume that there exists a continuous cdf Ψ with support in $[0,\infty)$ and mean $m_\Psi > 0$ such that

$$\bar{G}^*_{n,i}(t) \geq \bar{\Psi}(t), \quad \forall t, n, i, \tag{6}$$

or, by the definition of $G^*_{n,i}$,

$$\bar{G}_{n,i}(t) \geq \bar{\Psi}(\lambda_{n,i}t), \quad \forall t, n, i, \tag{7}$$

which, by (3), implies

$$\hat{G}_{n,i}(t) \geq m_\Psi \hat{\Psi}(\lambda_{n,i}t), \quad \forall t, n, i. \tag{8}$$

Let us elaborate a bit on the assumption in (6). Consider the L_1 distance between Ψ and $G^*_{n,i}$, which, by (6), is given by

$$\|G^*_{n,i} - \Psi\|_1 = \|\bar{G}^*_{n,i} - \bar{\Psi}\|_1 = \int_0^\infty [\bar{G}^*_{n,i}(t) - \bar{\Psi}(t)]dt = 1 - m_\Psi.$$

Since $\|G^*_{n,i} - \Psi\|_1 \geq 0$, it follows that $m_\Psi \leq 1$. Note that all $G^*_{n,i}$ live on the sphere of radius $1 - m_\Psi$ centered at Ψ. Since both $G^*_{n,i}$ and Ψ are continuous, $m_\Psi = 1$ if and only if $G^*_{n,i}(t) = \Psi(t)$ or, equivalently, if and only if $G_{n,i}(t) = \Psi(\lambda_{n,i}t)$ for all t, n and i. Intuitively, the function Ψ controls the variability within the family of cdfs $\mathcal{G} = \{G^*_{n,i} : n \geq i \geq 1\}$, and m_Ψ is a measure of this variability. When $m_\Psi \to 0$, the constraint (6) becomes empty, and $G^*_{n,i}$ could be very different from each other. As m_Ψ increases, $G^*_{n,i}$ become more and more similar to each other. When $m_\Psi = 1$, $G^*_{n,i}$ degenerates to a single distribution Ψ, in which case, $G_{n,i}$ are all from the scale family[3] as $G_{n,i}(t) = \Psi(\lambda_{n,i}t)$ from (5).

The most important example of the degenerate case $m_\Psi = 1$ is when all request processes are Poisson, i.e. $G_{n,i}(t) = 1 - e^{-\lambda_{n,i}t}$ with $\Psi(t) = 1 - e^{-t}$. Non-Poisson examples include Erlang distributions with the same number of stages, Gamma distributions with the same shape parameter, and Weibull distributions with the same shape parameter.

An important example of the non-degenerate case is when $G_{n,i}$ are from a finite number, J, of scale families, i.e. $\mathcal{G} = \{\Psi_1, \ldots, \Psi_J\}$ for some distinct cdfs Ψ_j with $m_{\Psi_j} = 1$. More specifically, let $\mathcal{P}_1, \ldots, \mathcal{P}_J$ be a partition of the set $\{(n,i) \in \mathbb{N}^2 : n \geq i \geq 1\}$ such that $G^*_{n,i} = \Psi_j$ for all $(n,i) \in \mathcal{P}_j$. Note that (6) holds with $\Psi(t) = \max_{1\leq j\leq J} \Psi_j(t)$ in this case. However, $m_\Psi < 1$ unless $J = 1$, which reduces to the degenerate case.

Let $N_n := \{t_n(k), k \in \mathbb{Z}\}$ be the point process resulting from the superposition of the n independent point processes $N_{n,1}, \ldots, N_{n,n}$, where $-\infty \leq \cdots < t_n(-1) < t_n(0) \leq 0 < t_n(1) < \cdots \leq \infty$. Note that we have used the fact that the points $t_n(k)$ are distinct with probability one [3, Property 1.1.1]. Let $\mathbb{P}^0_n$ be the Palm probability[4] associated with N_n, and $\mathbb{E}^0_n$ the associated expectation operator. Under $\mathbb{P}^0_n$ a content is requested at $t = 0$, i.e. $\mathbb{P}^0_n[t_n(0) = 0] = 1$. Let $X^0_n \in \{1, \ldots, n\}$ denote this content. It is known that (see e.g. [3, Section 1.4.2])

$$\mathbb{P}^0_n[X^0_n = i] = \frac{\lambda_{n,i}}{\Lambda_n} := p_{n,i}, \tag{9}$$

where $\Lambda_n := \sum_{i=1}^n \lambda_{n,i}$, and

$$\mathbb{P}^0_n[A] = \sum_{i=1}^n p_{n,i}\mathbb{P}^0_{n,i}[A] \tag{10}$$

for any event A.

[3]Recall that a family of cdfs $F(st)$, indexed by a *scale parameter* $s > 0$, is called the scale family with standard cdf F.

[4]Again, readers unfamiliar with Palm probability can think of $\mathbb{P}^0_n$ as being defined by $\mathbb{P}^0_n[A] = \mathbb{P}[A \mid t_n(0) = 0]$ for any event A, i.e. the conditional probability conditioned on the event that a request arrives at time 0.

2.2 Content Popularity

The probability $p_{n,i}$ defined in (9) gives the popularity of content i. Previous work (see e.g. [17] and references therein) shows that the popularity distribution $\{p_{n,1}, \ldots, p_{n,n}\}$ usually follows Zipf's law,

$$p_{n,i} = \frac{i^{-\alpha}}{\sum_{j=1}^{n} j^{-\alpha}}, \tag{11}$$

where $\alpha \geq 0$ and most often $\alpha \in (0,1)$. This will be the main example of popularity distribution used throughout the rest of the paper.

In [14], the popularity distribution is assumed to be given by

$$p_{n,i} = F\left(\frac{i}{n}\right) - F\left(\frac{i-1}{n}\right), \tag{12}$$

where F is a continuously differentiable cdf with support in $[0,1]$. With some slight modification, (12) can be extended to include (11) as a special case. Note that (12) does not assume the $p_{n,i}$'s are ordered in i.

In this paper, we consider more general popularity distributions, which include as special cases both (11) and (12) with the mild condition that $F' > 0$ a.e. on $[0,1]$. Let σ_i be the index of the i-th most popular content, i.e.

$$p_{n,\sigma_1} \geq p_{n,\sigma_2} \geq \cdots \geq p_{n,\sigma_n} \tag{13}$$

is the sequence $p_{n,1}, \ldots, p_{n,n}$ rearranged in decreasing order. Define the tail $\bar{P}_n$ of the content popularity distribution by

$$\bar{P}_n(i) = \sum_{k=i+1}^{n} p_{n,\sigma_k}, \tag{14}$$

which is the aggregate popularity of the $n-i$ least popular contents. Roughly speaking, we will focus on popularity distributions whose values $\bar{P}_n(i)$ are of the same order for i around C_n. This will be made more precise later; see assumption (P1) in Section 3.1.3.

2.3 TTL Approximation

Let $Y_{n,i}(t) = 1$ if content i is requested during the interval $[-t, 0)$ and $Y_{n,i}(t) = 0$ otherwise. With this notation,

$$Y_n(t) := \sum_{i=1}^{n} Y_{n,i}(t) \tag{15}$$

is the number of distinct contents requested during $[-t, 0)$. Let $[-\tau_n, 0)$ be the smallest past interval in which C_n distinct contents are referenced, i.e.,

$$\tau_n = \inf\{t : Y_n(t) \geq C_n\}. \tag{16}$$

Note that if we reverse the arrow of time, we obtain statistically the same request processes, and τ_n is a stopping time for the process $Y_n(t)$.

In an LRU cache, a content that is least recently referenced is evicted when another content needs to be added to the full cache. Thus a request for content i results in a cache hit if and only if i is among the C_n distinct most recently referenced contents. By stationarity, we can always assume that this request arrives at $t = 0$. Thus the stationary hit probability of an LRU cache is given by

$$H_n^{\mathrm{LRU}} = \mathbb{P}_n^0[Y_{n,X_n^0}(\tau_n) = 1]. \tag{17}$$

Similarly, the stationary hit probability of content i in an LRU cache is given by

$$H_{n,i}^{\mathrm{LRU}} = \mathbb{P}_{n,i}^0[Y_{n,i}(\tau_n) = 1], \tag{18}$$

By (10), H_n^{LRU} and $H_{n,i}^{\mathrm{LRU}}$ are related by

$$H_n^{\mathrm{LRU}} = \sum_{i=1}^{n} p_{n,i} H_{n,i}^{\mathrm{LRU}}. \tag{19}$$

In a TTL cache, when a content is added to the cache, its associated time-to-live timer is set to its maximum value T. The content is evicted from the cache when the timer expires. The capacity of the cache is assumed to be large enough to hold all contents with non-expired timers. In this paper, we consider the so-called TTL cache with reset, which always resets the associated timer to T when a cache hit occurs. Thus a request for content i results in a cache hit if and only if i is referenced in a past window of length T. The stationary hit probability is then given by

$$H_n^{\mathrm{TTL}}(T) = \mathbb{P}_n^0[Y_{n,X_n^0}(T) = 1], \tag{20}$$

and that for content i by

$$H_{n,i}^{\mathrm{TTL}}(T) = \mathbb{P}_{n,i}^0[Y_{n,i}(T) = 1], \tag{21}$$

which will be shown to equal $G_{n,i}(T)$ in Lemma 4.8. By (10), $H_n^{\mathrm{TTL}}(T)$ and $H_{n,i}^{\mathrm{TTL}}(T)$ are related by

$$H_n^{\mathrm{TTL}}(T) = \sum_{i=1}^{n} p_{n,i} H_{n,i}^{\mathrm{TTL}}(T). \tag{22}$$

The TTL approximation was first introduced by Fagin for IRM requests [14], later rediscovered for independent Poisson request processes [8] and extended to renewal request processes [18], in the latter two cases without theoretical basis. It should be noticed that Fagin's result can be reproduced [1] by restricting the support of the distribution to $[0, 1]$ in Theorem 4 in [23]. Also, Theorem 1 in [20] proves that the individual content hit probability in an LRU cache converges to the corresponding quantity in a TTL cache as the number of items increases to infinity, when contents are requested according to independent Poisson processes and when there is only a finite number of types of contents; see discussion after Example 4.5.

We now present it for general independent stationary and ergodic request processes. Let

$$K_n(T) := \mathbb{E}[Y_n(T)] \tag{23}$$

denote the expected number of contents in a TTL cache with timer value T, where Y_n is defined in (15). It will be shown in Lemma 4.9 that $K_n(T) = \sum_{i=1}^{n} \hat{G}_{n,i}(T)$. Given the size C_n of an LRU cache, let T_n satisfy

$$C_n = K_n(T_n) = \sum_{i=1}^{n} \hat{G}_{n,i}(T_n). \tag{24}$$

The time T_n is the *characteristic time* of the LRU cache. The TTL approximation then approximates the hit probabilities of the LRU cache by those of a TTL cache with timer value T_n, i.e.

$$H_{n,i}^{\mathrm{LRU}} \approx H_{n,i}^{\mathrm{TTL}}(T_n), \quad \forall i = 1, \ldots, n.$$

For Poisson requests, (24) takes the familiar form

$$C_n = \sum_{i=1}^{n} (1 - e^{-\lambda_{n,i} T_n}).$$

Note that the TTL approximation for general independent stationary and ergodic processes takes the same form as for renewal processes [18], which is not surprising in view of Theorem 2 in [19].

In Section 4, we show that, as C_n and n become large, the TTL approximation becomes exact, i.e. an LRU cache behaves like a TTL cache with a TTL approximation timer value equal to the LRU characteristic time.

3 OVERVIEW OF MAIN RESULTS

In this section we present the main results of the paper. Section 3.1 collects various assumptions used in the main results and discusses their relations. The mains results are presented in Section 3.2.

3.1 Assumptions

We divide the assumptions into three categories according to whether they concern cache size, request processes, or content popularity distribution.

3.1.1 Cache size. Throughout the paper, it is assumed that the cache size $C_n \in (0, n)$ and $C_n \to \infty$ as $n \to \infty$. In addition, each result assumes one of the following conditions.

(C1) $C_n \leq \beta_1 n$ for some $\beta_1 \in (0, m_\Psi)$ and n large enough, where m_Ψ is the mean of Ψ in (6).
(C2) $C_n \sim \beta_0 n$ for some $\beta_0 \in (0, 1)$.

Note that (C2) requires C_n to scale linearly in n while (C1) only requires C_n to scale at most linearly. For $\beta_0 < m_\Psi$, (C2) $\Longrightarrow$ (C1).

3.1.2 Request processes. The requests for different contents follow independent stationary and ergodic simple point processes. The request process for content i has continuous inter-request distribution satisfying (6). In addition, each result assumes one of the following conditions, with $\mathcal{G}_i := \{G^*_{n,i} : n \geq i\}$ and $\mathcal{G} := \bigcup_{i=1}^{\infty} \mathcal{G}_i$, where $G^*_{n,i}$ is defined in (5),

(R1) Given i, $\mathcal{G}_i$ is equicontinuous[5].
(R2) $\mathcal{G}$ is equicontinuous.
(R3) $|\mathcal{G}| < \infty$, i.e. the inter-request distributions are from a finite number of scale families.
(R4) $\mathcal{G} = \{\Psi\}$, i.e. the inter-request distributions are from a single scale family.
(R5) $\mathcal{G}$ is uniformly Lipschitz continuous[6].
(R6) There exist a constant B and $\rho \in (0, 1]$ such that

$$|G(t) - G(t \pm xt)| \leq Bx, \quad \text{for } x \in [0, \rho], \forall t \text{ and } \forall G \in \mathcal{G}. \tag{25}$$

By Lemma A.1, (R1) (resp. (R2)) holds if $\mathcal{G}_i$ (resp. $\mathcal{G}$) is composed of a finite family of continuous cdfs. Hence, (R4) $\Longrightarrow$ (R3) $\Longrightarrow$ (R2) $\Longrightarrow$ (R1). Note also that (R5) $\Longrightarrow$ (R2). Examples of (R5) include families of distributions that have densities with a common upper bound. The last condition (R6) can be thought of as some kind of uniform Lipschitz continuity, where the bound depends on the relative deviation of the arguments rather than on the absolute deviation as in (R5). Condition (R6) is satisfied if the inter-request distributions are all exponential, which corresponding to Poisson requests (Example 5.3), or, more generally, if (R3) holds with every $G \in \mathcal{G}$ having a continuous density (see Example 5.4, which also includes an example with infinite $\mathcal{G}$). Note that (R6) implies uniform Lipschitz continuity for t strictly bounded away from zero, which is in fact all we need when working with (R5), so for our purpose (R6) is stronger than (R5).

3.1.3 Popularity distribution. Each result assumes one of the following conditions for content popularity distribution.

[5]A family of functions $\mathcal{F}$ is *equicontinuous* if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $|x_1 - x_2| < \delta$ implies $|f(x_1) - f(x_2)| < \epsilon$ for every $f \in \mathcal{F}$. There is another commonly used definition of equicontinuity, which is a weaker notion in general but turns out to be equivalent to the former in our setting.
[6]A family of functions $\mathcal{F}$ is *uniformly Lipschitz continuous* if there exists an $M > 0$ such that $|f(x_1) - f(x_2)| < M|x_1 - x_2|$ for every x_1, x_2 and every $f \in \mathcal{F}$.

(P1) There exist constants $\kappa_1 \in (\frac{1}{m_\Psi}, \frac{1}{\beta_1})$ for β_1 in (C1), $\kappa_2 \in [0,1]$ and $\gamma \in (0,1)$ such that for all sufficiently large n, the tail popularity $\bar{P}_n$ defined in (14) satisfies

$$\bar{P}_n(\lceil \kappa_1 C_n \rceil) > \gamma \bar{P}_n(\lfloor \kappa_2 C_n \rfloor). \tag{26}$$

(P2) Fagin's condition: for some continuous function f defined on $(0,1]$ such that $f > 0$ a.e. and $\lim_{x\to 0+} f(x) \in [0,+\infty]$, and for some $z_{n,i} \in [\frac{i-1}{n}, \frac{i}{n}]$, the popularities $p_{n,i} \sim g_n f(z_{n,i})$ uniformly in i, i.e.

$$\max_{1\le i\le n} \left| \frac{g_n f(z_{n,i})}{p_{n,i}} - 1 \right| \to 0, \quad \text{as } n \to \infty. \tag{27}$$

(P3) The generalization (58) of (P2) from a single function f to a finite number of functions f_j's.

For a discussion of (P1), see Remark 1 after Proposition 4.4. Note that (P2) is slightly more general than Fagin's original condition (12). Note also (P2) $\Longrightarrow$ (P3) $\Longrightarrow$ (P1). Example 4.5 shows that the Zipfian popularity distribution in (11) with $\alpha \ge 0$ satisfies (P1). Example 6.2 shows that it also satisfies (P2) and hence (P3).

The common assumptions that $C_n \to \infty$ as $n \to \infty$ and that requests for different contents are described by mutually independent stationary and ergodic processes satisfying (6) will be assumed without explicit mentioning throughout the rest of the paper.

3.2 Main Results

In this section we present the main results of the paper. The first establishes that individual content hit probabilities under LRU converge to those under TTL as the cache size C_n and the number of contents n go to infinity, provided the timer values for all contents are set to the LRU characteristic time T_n introduced in the previous section, and provided the inter-request time distributions satisfy certain continuity properties.

Result 1 (Proposition 4.4). *Under assumptions* (C1), (R1) *and* (P1), *TTL approximation is asymptotically exact for content i, i.e.*

$$\left| H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n) \right| \to 0, \quad \text{as } n \to \infty.$$

Under assumptions (C1), (R2) *and* (P1), *TTL approximation is asymptotically exact uniformly for all contents, i.e.*

$$\max_{1\le i\le n} \left| H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n) \right| \to 0, \quad \text{as } n \to \infty.$$

The next result provides a uniform bound for the rates at which individual content hit probabilities under LRU converge to those under TTL under a slightly stronger Lipschitz continuity property.

Result 2 (Proposition 5.1). *Under assumptions* (C1), (R5) *and* (P1), *the following holds,*

$$\max_{1\le i\le n} \left| H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n) \right| = O\left(\left(\frac{\log C_n}{C_n} \right)^{\frac{1}{4}} \right).$$

The above rate of convergence is slow. This is improved in the next result where it is shown to be $O((\log C_n/C_n)^{1/2})$ under slightly stronger assumptions regarding the marginal inter-request time distributions. However, numerical results (see e.g [17]) suggest that the convergence rate might be faster than proved here.

Result 3 (Proposition 5.2). *Under assumptions* (C1), (R6) *and* (P1), *the following holds,*

$$\max_{1\le i\le n}\left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| = O\left(\sqrt{\frac{\log C_n}{C_n}}\right).$$

The last two results include extensions of Fagin's results for IRM to the case where content requests are described by mutually independent stationary and ergodic processes where the marginal inter-request time distributions satisfy mild continuity properties.

Result 4 (Proposition 6.3). *Under assumptions* (C2), (R4) *and* (P2), *the following holds,*

$$H_n^{\mathrm{LRU}} \to \int_0^1 f(x)\Psi(\nu_0 f(x))dx, \quad \text{as } n\to\infty,$$

where ν_0 *the unique real number in* $(0,\infty)$ *that satisfies*

$$\int_0^1 \hat{\Psi}(\nu_0 f(x))dx = \beta_0.$$

Result 4 considers a single class of contents in the sense that there is a single f and a single Ψ for all contents. The following result extends it to J classes of contents, where class j has a fraction b_j of the total contents, and each class j satisfies the assumptions in Result 4 with potentially different f_j and Ψ_j. See Proposition 6.4 for a more precise statement of (R3) and (P3).

Result 5 (Proposition 6.4). *Under assumptions* (C2), (R3) *and* (P3), *the following holds,*

$$H_n^{\mathrm{LRU}} \to \sum_{j=1}^{J} b_j \int_0^1 f_j(x)\Psi_j(\nu_0 f_j(x))dx, \quad \text{as } n\to\infty,$$

where ν_0 *the unique real number in* $(0,\infty)$ *that satisfies*

$$\sum_{j=1}^{J} b_j \int_0^1 \hat{\Psi}_j(\nu_0 f_j(x))dx = \beta_0.$$

4 ASYMPTOTIC EXACTNESS

It has been observed numerically in [17] that the TTL approximation is very accurate uniformly for contents of a wide range of popularity rank when the request processes are all Poisson. In this section, we prove that under some general conditions, the TTL approximation is exact in the large system regime, in the sense that individual content hit probabilities under LRU converge uniformly to those under TTL using the LRU characteristic time.

The following bounds on the LRU characteristic time T_n, which may be of interest in their own right, will be used in the proof of the main result, Proposition 4.4. The proof is found in Section 4.1.

Proposition 4.1. *The characteristic time* T_n *defined by* (24) *exists and is unique. For any* $n_1 \in (C_n/m_\Psi, n]$, *which exists if* $C_n < nm_\Psi$, *we have*

$$T_n \le \frac{\nu_0}{\lambda_{n,\sigma_{n_1}}}, \tag{28}$$

where σ_{n_1} *is defined in* (13), *and* ν_0, *which exists, is any constant that satisfies*

$$\hat{\Psi}(\nu_0) \ge \frac{C_n}{n_1 m_\Psi}.$$

For any $n_2 \le C_n$,

$$T_n \ge \frac{C_n - n_2}{\Lambda_n \bar{P}_n(n_2)}. \tag{29}$$

The following examples show that Proposition 4.1 yields the same scaling order of T_n as in [17, Eq. (7)] for Zipfian popularity distribution with $\alpha \neq 1$, but for request processes more general than Poisson.

Example 4.2. Consider Zipfian popularity distribution in (11) with $\alpha \in (0,1)$. In this case, we need $C_n = \Omega(n)$ so that the cache stores a nonnegligible fraction of the files in the sense that $P_n(C_n)$ does not vanish as n increases. Assume $C_n \sim \beta_0 n$ with $\beta_0 \in (0, m_\Psi)$. Setting $n_1 = n$ in (28), we obtain

$$T_n \le \frac{\nu_0}{p_{n,n}\Lambda_n} \sim \frac{\nu_0 n}{(1-\alpha)\Lambda_n},$$

where ν_0 satisfies $\hat{\Psi}(\nu_0) > \beta_0/m_\Psi$. Setting $n_2 = 0$ in (29), we obtain

$$T_n \ge \frac{C_n}{\Lambda_n} \sim \frac{\beta_0 n}{\Lambda_n}.$$

Note that

$$p_{n,\sigma_n} = p_{n,n} = \frac{n^{-\alpha}}{\sum_{j=1}^n j^{-\alpha}} \sim (1-\alpha)n^{-1},$$

where the last step follows from the well-known asymptotics (see e.g. [2, Theorem 3.2]) $\sum_{j=1}^n j^{-\alpha} \sim n^{1-\alpha}/(1-\alpha)$ for large n. Therefore, $T_n = \Theta\left(n\Lambda_n^{-1}\right)$. In particular, if $\lambda_{n,i} = i^{-\alpha}$, then $\Lambda_n \sim n^{1-\alpha}/(1-\alpha)$ and hence $T_n = \Theta(n^\alpha)$.

Example 4.3. Consider Zipfian popularity distribution in (11) with $\alpha > 1$. In this case, $P_n(C_n)$ never vanishes as long as $C_n \ge 1$. Assume $C_n \le \beta_0 n$ with $\beta_0 \in (0, m_\Psi)$. Consider the limit $C_n \to \infty$. Setting $n_1 \sim \kappa_1 C_n$ in (28) with $\kappa_1 \in (\frac{1}{m_\Psi}, \frac{1}{\beta_0})$, we obtain

$$p_{n,\sigma_{n_1}} = p_{n,n_1} = \frac{n_1^{-\alpha}}{\sum_{j=1}^n j^{-\alpha}} \sim \frac{1}{\kappa_2^\alpha C_n^\alpha \zeta(\alpha)},$$

and hence

$$T_n \le \frac{\nu_0 \kappa_1^\alpha \zeta(\alpha) C_n^\alpha}{\Lambda_n},$$

where ν_0 satisfies $\hat{\Psi}(\nu_0) > (\kappa_1 m_\Psi)^{-1}$. Setting $n_2 \sim \kappa_2 C_n$ in (29) with $\kappa_2 \in (0,1)$, we obtain

$$\bar{P}_n(n_2) \sim \frac{n_2^{1-\alpha}}{(1-\alpha)\zeta(\alpha)},$$

where $\zeta(\alpha) = \sum_{j=1}^\infty j^{-\alpha}$ is the Riemann zeta function. Thus

$$T_n \ge \frac{C_n - n_2}{\Lambda_n \bar{P}(n_2)} \sim \frac{(1-\alpha)\zeta(\alpha)(C_n - n_2)}{\Lambda_n n_2^{1-\alpha}} \sim \frac{(1-\alpha)\zeta(\alpha)(1-\kappa_2)C_n^\alpha}{\Lambda_n \kappa_2^{1-\alpha}},$$

Therefore, $T_n = \Theta\left(C_n^\alpha \Lambda_n^{-1}\right)$. In particular, if $C_n = \Theta(n)$ and $\lambda_i = i^{-\alpha}$, then $\Lambda_n \sim \zeta(\alpha)$ and hence $T_n = \Theta(n^\alpha)$. However, we do not need to have C_n scale linearly in n.

Proposition 4.4 is the main result, which provides sufficient conditions for the hit probabilities in the TTL approximation to converge to the corresponding hit probabilities in the LRU cache. The proof is found in Section 4.2.

Proposition 4.4. *Under assumptions (C1), (R1) and (P1), TTL approximation is asymptotically exact for content i, i.e.*

$$\left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| \to 0, \quad \text{as } n \to \infty. \tag{30}$$

Under assumptions (C1), (R2) and (P1), TTL approximation is asymptotically exact uniformly for all contents, i.e.

$$\max_{1\le i\le n}\left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| \to 0, \quad \text{as } n \to \infty. \tag{31}$$

Remark 1. Condition (P1) requires that the popularity distribution $\bar{P}_n(i)$ take values of the same order for i around C_n, as alluded to in Section 2.2. Intuitively, this means $\bar{P}_n(i)$ should not change abruptly around $i = C_n$. In a stronger form obtained by setting $\kappa_2 = 0$, (26) reads $\bar{P}_n(\lceil \kappa_1 C_n \rceil) > \gamma$, which means that even with a slightly larger cache, the contents that cannot fit into the cache have an aggregate probability at least γ, or equivalently, the optimal static caching policy has a miss probability at least γ. For Zipfian popularity in (11), this stronger form is satisfied only for $\alpha \le 1$, while (26) is satisfied for all $\alpha \ge 0$ as shown in Example 4.5.

Example 4.5. Consider the Zipfian popularity distribution in (11). We first check that assumption (P1) is satisfied for all $\alpha \ge 0$. For large n,

$$\bar{P}_n(i) \sim \begin{cases} \frac{n^{1-\alpha}-i^{1-\alpha}}{n^{1-\alpha}}, & \text{if } 0 \le \alpha < 1; \\ \frac{\log n - \log i}{\log n}, & \text{if } \alpha = 1; \\ \frac{i^{1-\alpha}}{(1-\alpha)\zeta(\alpha)}, & \text{if } \alpha > 1. \end{cases}$$

Thus

$$\liminf_{n\to\infty} \frac{\bar{P}_n(\lceil \kappa_1 C_n \rceil)}{\bar{P}_n(\lfloor \kappa_2 C_n \rfloor)} \ge \begin{cases} 1 - (\kappa_1\beta_1)^{1-\alpha}, & \text{if } 0 \le \alpha < 1; \\ 1, & \text{if } \alpha = 1; \\ \left(\frac{\kappa_2}{\kappa_1}\right)^{\alpha-1}, & \text{if } \alpha > 1. \end{cases}$$

In all cases, the above guarantees the existence of a $\gamma \in (0,1)$ for which (26) holds. Note that for $\alpha \le 1$, we can set $\kappa_2 = 0$. If $m_\Psi = 1$, then $\mathcal{G} = \{\Psi\}$, which satisfies (R2) by Lemma A.1. Thus (31) holds for any C_n satisfying (C1). In particular, (31) holds when all request processes are Poisson.

As indicated in Section 2.3, Hirade and Osogami proved in [20] that the individual content hit probability in an LRU cache converges to the individual content hit probability in a TTL cache for Poisson requests as the number of contents increases to infinity. More precisely, they consider nN contents, $e_{i,j}$, $i = 1, \ldots, N$, $j = 1, \ldots, n$, each of size $1/n$, where successive requests for content $e_{i,j}$ follow a Poisson process with rate λ_i. These Poisson processes are assumed to be mutually independent. Note that in this setting there is only a finite number of types of requests ($= N$)[7]. Define $F_i(t) = 1 - \exp(-\lambda_i t)$. It is shown in [20, Theorem 1] that the probability, $p_{i,j}^{(n)}$, that content $e_{i,j}$ is in an LRU cache converges to $F_i(T)$ as $n \to \infty$, where T is the unique solution of the equation $\sum_{i=1}^{N} F_i(T) = K$, with $K < N$ being the size of the cache. By performing the substitutions $n \to nN$, $G_{nN,(n-1)i+j}(\cdot) \to F_i(\cdot)$ for $j = 1, \ldots, n$, $i = 1, \ldots, N$ and $C_n \to nK$ (with these substitutions the ratio "cache size/content size = nK" is the same as in [20]), we get from (30),

$$H_{nN,i}^{\mathrm{LRU}} \sim H_{nN,i}^{\mathrm{TTL}}(T_{nN}) = F_i(T_{nN}) \quad \text{as } n \to \infty,$$

[7] This can be considered as a special case of the setting in Proposition 6.4 with N classes, each consisting of n equally popular contents. However, Theorem 1 of [20] concerns hit probabilities of individual contents, while Proposition 6.4 concerns average hit probability.

where (see (24)) T_{nN} is the unique t satisfying the equation $Kn = \sum_{i=1}^{nN} \hat{G}_{nN,i}(t) = \sum_{i=1}^{N} nF_i(t)$, or equivalently, $K = \sum_{i=1}^{N} F_i(t)$. We now check the conditions (C1), (R1) and (P1) for (30). Condition (C1) reads $nK \le \beta_1 nN$, which holds for any $K/N \le \beta_1 < 1$ (note that $m_\Psi = 1$ since requests are Poisson). By Lemma A.1, $\mathcal{G}_i = \{\Psi\}$ with $\Psi(t) = 1 - e^{-t}$ is equicontinuous, satisfying (R1). To check (P1), we first observe that contents $e_{i,1}, \ldots, e_{i,n}$ have the same popularity $r_i/n \in (0,1)$ with $\sum_{i=1}^{N} r_i = 1$. Hence, $\bar{P}_{nN}(\lceil \kappa_1 C_n \rceil) \gtrsim N(1-\kappa_1\beta_1)\min_{1\le i\le N} r_i := \gamma$. Since one can find $\kappa_1 \in (1, 1/\beta_1)$ such that $\gamma \in (0,1)$, we have shown that (26) holds with this γ and $\kappa_2 = 0$.

Note that a similar fluid approximation for an LRU cache is developed in [27], which considers dependent and so-called time-asymptotically stationary requests. However, the modification introduced to deal with the dependence structure renders the new approximation unsuitable for a re-interpretation as above. Thus the results therein do not apply to TTL approximations. Observe also that there is only empirical evidence but no theoretical proof that the fluid limit is an accurate approximation of the original LRU cache.

The following corollary considers the convergence of the aggregate hit probability.

Corollary 4.6. *Assume* (C1) *and* (P1). *Then as* $n \to \infty$,

$$\left|H_n^{\mathrm{LRU}} - H_n^{\mathrm{TTL}}(T_n)\right| \to 0, \tag{32}$$

if either (R2) *holds, or for each* i, (R1) *and the following hold*

$$\lim_{m\to\infty} \limsup_{n\to\infty} \bar{P}_n(m) = 0. \tag{33}$$

Proof. By (19) and (22), for any m,

$$\left|H_n^{\mathrm{LRU}} - H_n^{\mathrm{TTL}}(T_n)\right| \le \max_{1\le i\le m} \left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| + \bar{P}_n(m).$$

Suppose (R2) holds. Let $m = n$. Since $\bar{P}_n(n) = 0$, (32) follows from (31).

Suppose for each i, (R1) and (33) hold. Fix m and let $n \to \infty$. By (30),

$$\limsup_{n\to\infty} \left|H_n^{\mathrm{LRU}} - H_n^{\mathrm{TTL}}(T_n)\right| \le \limsup_{n\to\infty} \bar{P}_n(m).$$

Now let $m \to \infty$ and (32) follows (33). □

Example 4.7. For the Zipfian popularity distribution in (11),

$$\limsup_{n\to\infty} \bar{P}_n(m) = \begin{cases} 1, & \text{if } 0 \le \alpha \le 1; \\ \frac{m^{1-\alpha}}{(1-\alpha)\zeta(\alpha)}, & \text{if } \alpha > 1. \end{cases}$$

Thus (33) holds for $\alpha > 1$ but fails for $\alpha \in [0,1]$. For each i, if the standardized cdf $G_{n,i}^*$ is the same for all n, then $\mathcal{G}_i$ is a singleton and hence equicontinuous by Lemma A.1. In this case, (32) holds for $\alpha > 1$, but we cannot conclude the same for $\alpha \le 1$ without further assuming that $\mathcal{G}$ is equicontinuous. When $m_\Psi = 1$, in particular, when all request processes are Poisson, (32) holds. For Poisson requests, Fagin [14] has established the convergence for $\alpha \in (0,1)$ and $C_n \sim \beta_0 n$. We now see this is also true for $\alpha \ge 1$ and for C_n scaling sublinearly in n.

4.1 Proof of Proposition 4.1

We need the following two simple lemmas.

Lemma 4.8. *For* $i, j = 1, \ldots, n$, *and* $t > 0$,

$$\mathbb{P}_{n,j}^0[Y_{n,i}(t) = 1] = \mathbb{1}_{\{j=i\}} G_{n,i}(t) + \mathbb{1}_{\{j\ne i\}} \hat{G}_{n,i}(t). \tag{34}$$

Proof. For $i = j$, since $t_{n,i}(0) = 0$ a.s. under $\mathbb{P}^0_{n,i}$, we have

$$\mathbb{P}^0_{n,i}[Y_{n,i}(t) = 1] = \mathbb{P}^0_{n,i}[-t_{n,i}(-1) \le t] = G_{n,i}(t).$$

For $i \neq j$, the independence of the point processes $N_{n,i}$ and $N_{n,j}$ yields

$$\mathbb{P}^0_{n,j}[Y_{n,i}(t) = 1] = \mathbb{P}[Y_{n,i}(t) = 1];$$

see [3, Eq. (1.4.5)] for a more formal statement. Since $t_{n,i}(0) < 0$ a.s. under $\mathbb{P}$, we obtain

$$\mathbb{P}^0_{n,j}[Y_{n,i}(t) = 1] = \mathbb{P}[Y_{n,i}(t) = 1] = \mathbb{P}[-t_{n,i}(0) \le t] = \hat{G}_{n,i}(t), \tag{35}$$

where the last equality follows from (4). This completes the proof of (34). □

Lemma 4.9. *The function K_n defined in (23) satisfies the following,*

$$K_n(T) = \sum_{i=1}^{n} \hat{G}_{n,i}(T), \tag{36}$$

$$K_n'(T) = \sum_{i=1}^{n} \lambda_{n,i} \bar{G}_{n,i}(T). \tag{37}$$

The function K_n is concave on $[0, \infty)$ and strictly increasing at all $T \in [0, \infty)$ such that $K_n(T) < n$.

Proof. Using (23), (15) and (35), we obtain

$$K_n(T) = \mathbb{E}[Y_n(T)] = \sum_{i=1}^{n} \mathbb{P}[Y_{n,i}(T) = 1] = \sum_{i=1}^{n} \hat{G}_{n,i}(T),$$

proving (36). Taking the derivative of (36) w.r.t. T and using (4) yield (37). Note that K_n' is a decreasing function of T, from which it follows that $K_n(T)$ is concave.

Now we show that $K_n'(T) > 0$ at all T such that $K_n(T) < n$, from which it will follow that K_n is strictly increasing at all such T. Clearly $K_n'(T) \ge 0$ from (37). Assume that $K_n'(T) = 0$ for some $T > 0$. Then, $\bar{G}_{n,i}(T) = 0$ for all i, which, by monotonicity of $\bar{G}_{n,i}$, yields $\bar{G}_{n,i}(y) = 0$ for all $y \ge T$. Thus, by (4),

$$1 - \hat{G}_{n,i}(T) = \lambda_i \int_T^\infty \bar{G}_{n,i}(y)dy = 0,$$

which implies $K_n(T) = n$ by (36). Therefore, $K_n'(T) > 0$ for all T such that $K_n(T) < n$. □

Now we prove Proposition 4.1.

Proof of Proposition 4.1. The existence of T_n follows from the continuity of K_n, the facts $K_n(0) = 0$ and $\lim_{T\to\infty} K_n(T) = n$, and the Intermediate Value Theorem. Uniqueness follows from the strict monotonicity of K_n given by Lemma 4.9.

By (4) and the fact $\bar{G}_{n,i}(y) \le 1$, we have

$$\hat{G}_{n,i}(T_n) = \lambda_{n,i} \int_0^{T_n} \bar{G}_{n,i}(y)dy \le \lambda_{n,i} T_n.$$

Thus

$$C_n = K_n(T_n) = \sum_{i=1}^{n} \hat{G}_{n,i}(T_n) \le \sum_{i=1}^{n} \min\{1, \lambda_{n,i} T_n\} \le \sum_{i=1}^{n_2} 1 + \sum_{i=n_2+1}^{n} \lambda_{n,i} T_n = n_2 + \Lambda_n T_n \bar{P}_n(n_2),$$

from which (29) follows.

To prove (28), note that

$$C_n = \sum_{i=1}^{n} \hat{G}_{n,i}(T_n) \geq m_\Psi \sum_{i=1}^{n} \hat{\Psi}(\lambda_{n,i} T_n) \geq n_1 m_\Psi \hat{\Psi}(\lambda_{n,\sigma_{n_1}} T_n),$$

where the first inequality follows from (8), and the second from (9), (13), and the monotonicity of $\hat{\Psi}$. Since $C_n/(n_1 m_\Psi) < 1$ and $\hat{\Psi}$ is a continuous cdf, there exists a v_0 such that

$$\hat{\Psi}(v_0) \geq \frac{C_n}{n_1 m_\Psi}.$$

For any such v_0,

$$\hat{\Psi}(\lambda_{n,\sigma_{n_1}} T_n) \leq \frac{C_n}{n_1 m_\Psi} \leq \hat{\Psi}(v_0),$$

which, together with the monotonicity of $\hat{\Psi}$, yields (28). □

4.2 Proof of Proposition 4.4

The proof of Proposition 4.4 relies on the four lemmas below.

Note by (37) that $K'_n(T)$ is the aggregate miss rate of a TTL cache with timer T, and

$$\mu_n(T) = \frac{K'_n(T)}{\Lambda_n} = \sum_{i=1}^{n} p_{n,i} \bar{G}_{n,i}(T) \tag{38}$$

is the aggregate miss probability.

Lemma 4.10. *Assume (C1) and (P1). Then there exist strictly positive constants x_0, ϕ that do not depend on n, such that for $T \leq (1 + x_0)T_n$ and sufficiently large n,*

$$\mu_n(T) \geq \frac{\phi C_n}{\Lambda_n T_n}. \tag{39}$$

Proof. Recall the definition of κ_1 and κ_2 in the statement of Proposition 4.4. Let $n_1 = \lceil \kappa_1 C_n \rceil$, $n_2 = \lfloor \kappa_2 C_n \rfloor$. As $C_n/(n m_\Psi) \leq \beta_1/m_\Psi < 1$ for sufficiently large n by (C1) and $\hat{\Psi}$ is a continuous cdf with $\hat{\Psi}(0) = 0$, there exist v_0 and $x_0 > 0$ such that

$$1 > \hat{\Psi}((1 + x_0)v_0) \geq \hat{\Psi}(v_0) \geq \beta_1/m_\Psi \geq C_n/(n m_\Psi) \tag{40}$$

for sufficiently large n. Recall the content ordering (13). For sufficiently large n,

$$\begin{aligned}
\mu_n(T) &= \sum_{i=1}^{n} p_{n,\sigma_i} \bar{G}_{n,\sigma_i}(T) \\
&\geq \sum_{i=1}^{n} p_{n,\sigma_i} \bar{\Psi}(\lambda_{n,\sigma_i} T) && \text{by (7)} \\
&\geq \sum_{i=n_1+1}^{n} p_{n,\sigma_i} \bar{\Psi}(\lambda_{n,\sigma_i} T) \\
&\geq \bar{\Psi}(\lambda_{n,\sigma_{n_1}} T) \sum_{i=n_1+1}^{n} p_{n,\sigma_i} && \text{by (13)} \\
&= \bar{\Psi}(\lambda_{n,\sigma_{n_1}} T) \bar{P}_n(n_1).
\end{aligned} \tag{41}$$

Since $\mu_n(T)$ is monotonically decreasing in T, we obtain for $T \leq (1 + x_0)T_n$ and all sufficiently large n,

$$\mu_n(T) \geq \mu_n((1 + x_0)T_n)$$

$$
\begin{aligned}
&\geq \bar{\Psi}((1+x_0)\lambda_{n,\sigma_{n_1}} T_n)\bar{P}_n(n_1) && \text{by (41)}\\
&\geq \bar{\Psi}((1+x_0)\nu_0)\bar{P}_n(n_1) && \text{by (28)}\\
&\geq \frac{C_n - n_2}{\Lambda_n T_n}\bar{\Psi}((1+x_0)\nu_0)\frac{\bar{P}_n(n_1)}{\bar{P}_n(n_2)} && \text{by (29)}\\
&\geq \frac{(1-\kappa_2)C_n}{\Lambda_n T_n}\bar{\Psi}((1+x_0)\nu_0)\frac{\bar{P}_n(n_1)}{\bar{P}_n(n_2)}\\
&\geq \frac{(1-\kappa_2)C_n}{\Lambda_n T_n}\bar{\Psi}((1+x_0)\nu_0)\gamma && \text{by (26).}
\end{aligned}
$$

The last inequality yields (39) with $\phi = (1-\kappa_2)\gamma\bar{\Psi}((1+x_0)\nu_0)$ if $\bar{\Psi}((1+x_0)\nu_0) > 0$. Assume that $\bar{\Psi}((1+x_0)\nu_0) = 0$. This would imply that $\Psi(x) = 1$ for all $x \geq (1+x_0)\nu_0$ by monotonicity of Ψ, which would in turn imply that $1 - \hat{\Psi}((1+x_0)\nu_0) = (1/m_\Psi)\int_{(1+x_0)\nu_0}^{\infty}\bar{\Psi}(t)dt = 0$, contradicting (40). Therefore, we indeed have $\bar{\Psi}((1+x_0)\nu_0) > 0$, which completes the proof. □

Lemma 4.11 (Kolmogorov's inequality [26, Section 19.1]). *Let $X_1, \ldots, X_n$ be independent random variables such that $\mathbb{E}X_i = 0$ and $|X_i| \leq b$ for all i. Then for any $x > 0$,*

$$
\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq x\right] \leq \exp\left\{-\frac{x^2}{4\max\{s_n^2, bx\}}\right\}, \tag{42}
$$

where $s_n^2 = \sum_{i=1}^{n}\mathbb{E}X_i^2$ is the variance of $\sum_{i=1}^{n} X_i$.

The next lemma shows that τ_n is concentrated around T_n.

Lemma 4.12. *Assume (39) holds for $T \leq (1+x_0)T_n$. Then for $0 \leq x \leq \min\{1, x_0\}$,*

$$
\mathbb{P}^0_{n,i}[\tau_n > (1+x)T_n] \leq \exp\left\{-\frac{(\phi x C_n)^2}{4(1+x)C_n + 4}\right\}.
$$

If, in addition, $\phi x C_n \geq 1$, then

$$
\mathbb{P}^0_{n,i}[\tau_n < (1-x)T_n] \leq \exp\left\{-\frac{(\phi x C_n - 1)^2}{4C_n + 4}\right\}.
$$

Proof. Let $T_n^+ = (1+x)T_n$ and $T_n^- = (1-x)T_n$. Note that

$$
K_n(T_n^+) - C_n = K_n(T_n^+) - K_n(T_n) = \int_{T_n}^{T_n^+} K_n'(T)dT,
$$

which, by (38) and (39), yields

$$
K_n(T_n^+) - C_n \geq \int_{T_n}^{T_n^+} \frac{\phi C_n}{T_n} dT = \phi x C_n. \tag{43}
$$

Since $T_n = (T_n^+ + T_n^-)/2$, the concavity of K_n yields

$$
C_n - K_n(T_n^-) = K_n(T_n) - K_n(T_n^-) \geq K_n(T_n^+) - K_n(T_n) = K_n(T_n^+) - C_n \geq \phi x C_n \tag{44}
$$

by (43). Note that by (15), (34) and (36), we have

$$
\mathbb{E}^0_{n,i}[Y_n(T)] = \sum_{j=1}^{n}\mathbb{E}^0_{n,i}[Y_{n,j}(T)] = K_n(T) + G_{n,i}(T) - \hat{G}_{n,i}(T).
$$

Since $G_{n,i}$ and $\hat{G}_{n,i}$ are both cdfs, we obtain

$$
K_n(T) - 1 \leq \mathbb{E}^0_{n,i}[Y_n(T)] \leq K_n(T) + 1. \tag{45}
$$

Using the definition of τ_n in (16), we obtain

$$\mathbb{P}^0_{n,i}[\tau_n > T_n^+] = \mathbb{P}^0_{n,i}[Y_n(T_n^+) \le C_n - 1] = \mathbb{P}^0_{n,i}\left\{Y_n(T_n^+) - \mathbb{E}^0_{n,i}[Y_n(T_n^+)] \le C_n - 1 - \mathbb{E}^0_{n,i}[Y_n(T_n^+)]\right\}.$$

By (45) and (43),

$$C_n - 1 - \mathbb{E}^0_{n,i}[Y_n(T_n^+)] \le C_n - K_n(T_n^+) \le -\phi x C_n.$$

Thus

$$\mathbb{P}^0_{n,i}[\tau_n > T_n^+] \le \mathbb{P}^0_{n,i}\left[Y_n(T_n^+) - \mathbb{E}^0_{n,i}[Y_n(T_n^+)] \le -\phi x C_n\right]. \tag{46}$$

Since the request processes $N_{n,1}, N_{n,2} \ldots, N_{n,n}$ are independent, so are the Bernoulli random variables $Y_{n,1}(t), Y_{n,2}(t), \ldots, Y_{n,n}(t)$ under $\mathbb{P}^0_{n,i}$. Thus

$$\begin{aligned}
\mathrm{var}^0_{n,i}[Y_n(T_n^+)] &= \sum_{j=1}^n \mathrm{var}^0_{n,i}[Y_{n,j}(T_n^+)] \le \sum_{j=1}^n \mathbb{E}^0_{n,i}[Y_{n,j}(T_n^+)] = \mathbb{E}^0_{n,i}[Y_n(T_n^+)] \\
&\le K_n(T_n^+) + 1 \quad \text{by (45)} \\
&\le (1+x)C_n + 1, \qquad (47)
\end{aligned}$$

where last step follows from the following consequence of the concavity of K_n

$$\frac{x}{1+x}K_n(0) + \frac{1}{1+x}K_n(T_n^+) \le K_n\left(\frac{T_n^+}{1+x}\right) = K_n(T_n) = C_n$$

and the fact $K_n(0) = 0$.

Note that $\left|Y_{n,i}(T) - \mathbb{E}^0_{n,i}[Y_{n,i}]\right| \le 1$. By applying Kolmogorov's inequality (42) with $b = 1$ and $s_n^2 \le (1+x)C_n + 1$ to the r.h.s. of (46), we obtain

$$\mathbb{P}^0_{n,i}[\tau_n > T_n^+] \le \exp\left\{-\frac{(\phi x C_n)^2}{4(1+x)C_n + 4}\right\}.$$

Similarly, if $\phi x C_n \ge 1$, we have

$$\begin{aligned}
\mathbb{P}^0_{n,i}[\tau_n < T_n^-] &= \mathbb{P}^0_{n,i}[Y_n(T_n^-) \ge C_n] \\
&= \mathbb{P}^0_{n,i}\left[Y_n(T_n^-) - \mathbb{E}^0_{n,i}[Y_n(T_n^-)] \ge C_n - \mathbb{E}^0_{n,i}[Y_n(T_n^-)]\right] \\
&\le \mathbb{P}^0_{n,i}\left[Y_n(T_n^-) - \mathbb{E}^0_{n,i}[Y_n(T_n^-)] \ge C_n - K_n(T_n^-) - 1\right] \quad \text{by (45)} \\
&\le \mathbb{P}^0_{n,i}\left[Y_n(T_n^-) - \mathbb{E}^0_{n,i}[Y_n(T_n^-)] \ge \phi x C_n - 1\right] \quad \text{by (44)} \\
&\le \exp\left\{-\frac{(\phi x C_n - 1)^2}{4C_n + 4}\right\} \quad \text{by (42)}.
\end{aligned}$$

□

Lemma 4.13.

$$\mathbb{P}^0_{n,i}[Y_{n,i}(\tau_n) = 1, \tau_n \le T] \le \mathbb{P}^0_{n,i}[Y_{n,i}(T) = 1, \tau_n \le T],$$

and

$$\mathbb{P}^0_{n,i}[Y_{n,i}(\tau_n) = 1, \tau_n \ge T] \ge \mathbb{P}^0_{n,i}[Y_{n,i}(T) = 1, \tau_n \ge T].$$

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

Now we prove Proposition 4.4.

Proof of Proposition 4.4. Fix an arbitrary $\epsilon > 0$. We show that for large enough n,

$$\left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| \le 2\epsilon. \tag{48}$$

The proof consists of two steps. We first show that $H_{n,i}^{\mathrm{TTL}}(T_n)$ is within ϵ distance from both $H_{n,i}^{\mathrm{TTL}}(T_n^+)$ and $H_{n,i}^{\mathrm{TTL}}(T_n^-)$ for some T_n^+ and T_n^- to be defined below. We then show that $H_{n,i}^{\mathrm{LRU}}$ is within ϵ distance from at least one of $H_{n,i}^{\mathrm{TTL}}(T_n^+)$ and $H_{n,i}^{\mathrm{TTL}}(T_n^-)$.

Let x_0 and ϕ be given by Lemma 4.10. Since the family $\mathcal{G}_i$ is equicontinuous by (R1), there exists $\xi_i(\epsilon) > 0$ such that $|t_1 - t_2| \le \xi_i(\epsilon)$ implies $|G_{n,i}^*(t_1) - G_{n,i}^*(t_2)| \le \epsilon$. Since $C_n \to \infty$ as $n \to \infty$, let n be sufficiently large so that

$$C_n \ge \max\left\{\frac{1}{\phi x_0}, \frac{1+\epsilon\xi_i(\epsilon)}{\phi\epsilon\xi_i(\epsilon)}\right\},$$

which guarantees the existence of an x satisfying the following,

$$\frac{1}{\phi C_n} \le x \le \min\left\{x_0, \frac{\epsilon\xi_i(\epsilon)}{1+\epsilon\xi_i(\epsilon)}\right\}. \tag{49}$$

Fix such an x. Let $T_n^+ = (1+x)T_n$, $T_n^- = (1-x)T_n$.

We first show

$$H_{n,i}^{\mathrm{TTL}}(T_n) - H_{n,i}^{\mathrm{TTL}}(T_n^-) \le \epsilon, \tag{50}$$

and

$$H_{n,i}^{\mathrm{TTL}}(T_n^+) - H_{n,i}^{\mathrm{TTL}}(T_n) \le \epsilon. \tag{51}$$

We only shown (50), as (51) follows from the same argument. By Lemma 4.8, (50) is the same as $G_{n,i}(T_n) - G_{n,i}(T_n^-) \le \epsilon$. Note that (this result holds regardless of the values of ϵ, $\xi_i(\epsilon)$ and $\lambda_{n,i}T_n$)

$$\max\left\{1 - \frac{1}{\epsilon\lambda_{n,i}T_n}, \frac{\xi_i(\epsilon)}{\lambda_{n,i}T_n}\right\} \ge \frac{\epsilon\xi_i(\epsilon)}{1+\epsilon\xi_i(\epsilon)}.$$

Since x satisfies (49), there are two cases: either $x \le \xi_i(\epsilon)/(\lambda_{n,i}T_n)$ or $x \le 1 - (\epsilon\lambda_{n,i}T_n)^{-1}$. In the first case, $|\lambda_{n,i}T_n - \lambda_{n,i}T_n^-| = x\lambda_{n,i}T_n \le \xi_i(\epsilon)$. Since $G_{n,i}(t) = G_{n,i}^*(\lambda_{n,i}t)$, using the definition of $\xi_i(\epsilon)$, we obtain $G_{n,i}(T_n) - G_{n,i}(T_n^-) \le \epsilon$. In the second case, note that

$$G_{n,i}(T_n) - G_{n,i}(T_n^-) \le 1 - G_{n,i}(T_n^-) = \bar{G}_{n,i}(T_n^-),$$

and

$$1/\lambda_{n,i} = \int_0^\infty \bar{G}_{n,i}(y)dy \ge \int_0^{T_n^-} \bar{G}_{n,i}(y)dy \ge T_n^-\bar{G}_{n,i}(T_n^-).$$

Thus

$$G_{n,i}(T_n) - G_{n,i}(T_n^-) \le \bar{G}_{n,i}(T_n^-) \le \frac{1}{\lambda_{n,i}T_n^-} \le \epsilon,$$

where the last inequality follows from the definition $T_n^- = (1-x)T_n$ and the condition $x \le 1 - (\epsilon\lambda_{n,i}T_n)^{-1}$. This proves (50).

Next we show (48). By Lemma 4.12, for sufficiently large C_n,

$$\left.\begin{array}{l}\mathbb{P}_{n,i}^0[\tau_n > T_n^+]\\ \mathbb{P}_{n,i}^0[\tau_n < T_n^-]\end{array}\right\} \le \exp\left\{-\frac{(\phi x C_n - 1)^2}{4(1+x)C_n + 4}\right\} \le \epsilon. \tag{52}$$

Note that

$$\begin{aligned} H_{n,i}^{\mathrm{LRU}} &= \mathbb{P}_{n,i}^0[Y_{n,i}(\tau_n) = 1] \\ &\ge \mathbb{P}_{n,i}^0[Y_{n,i}(\tau_n) = 1, \tau_n \ge T_n^-] \\ &\ge \mathbb{P}_{n,i}^0[Y_{n,i}(T_n^-) = 1, \tau_n \ge T_n^-] \quad \text{by Lemma 4.13} \end{aligned}$$

$$\geq \mathbb{P}^0_{n,i}[Y_{n,i}(T_n^-) = 1] - \mathbb{P}^0_{n,i}[\tau_n < T_n^-]$$
$$= H^{\mathrm{TTL}}_{n,i}(T_n^-) - \mathbb{P}^0_{n,i}[\tau_n < T_n^-],$$

which, by (50) and (52), yields

$$H^{\mathrm{TTL}}_{n,i}(T_n) - H^{\mathrm{LRU}}_{n,i} \leq H^{\mathrm{TTL}}_{n,i}(T_n) - H^{\mathrm{TTL}}_{n,i}(T_n^-) + \mathbb{P}^0_{n,i}[\tau_n < T_n^-] \leq 2\epsilon.$$

Note that similar bounds have been used for the shot noise model in [25].

For the other direction, note that

$$\begin{aligned} H^{\mathrm{LRU}}_{n,i} &\leq \mathbb{P}^0_{n,i}[Y_{n,i}(\tau_n) = 1, \tau_n \leq T_n^+] + \mathbb{P}^0_{n,i}[\tau_n > T_n^+] \\ &\leq \mathbb{P}^0_{n,i}[Y_{n,i}(T_n^+) = 1, \tau_n \leq T_n^+] + \mathbb{P}^0_{n,i}[\tau_n > T_n^+] \quad \text{by Lemma 4.13} \\ &\leq \mathbb{P}^0_{n,i}[Y_{n,i}(T_n^+) = 1] + \mathbb{P}^0_{n,i}[\tau_n > T_n^+] \\ &= H^{\mathrm{TTL}}_{n,i}(T_n^+) + \mathbb{P}^0_{n,i}[\tau_n > T_n^+], \end{aligned}$$

which, by (51) and (52), yields

$$H^{\mathrm{LRU}}_{n,i} - H^{\mathrm{TTL}}_{n,i}(T_n) \leq H^{\mathrm{TTL}}_{n,i}(T_n^+) - H^{\mathrm{TTL}}_{n,i}(T_n) + \mathbb{P}^0_{n,i}[\tau_n < T_n^+] \leq 2\epsilon.$$

Therefore, (48) holds, which proves (30).

Finally, (31) follows from the same argument with $\xi_i(\epsilon)$ replaced by $\xi(\epsilon)$, whose existence is guaranteed by (R2), i.e. the equicontinuity of the family $\mathcal{G}$. □

Remark 2. In the above proof of Proposition 4.4, the conditions (C1) and (P1) are used only to establish (39) in Lemma 4.10. Therefore, Proposition 4.4 and Corollary 4.6 will still hold if (C1) and (P1) are replaced by (39) or other conditions that imply (39).

Remark 3. Note that [17] provides a more concise argument to justify the TTL approximation in the case of Poisson requests, but the argument does not constitute a rigorous proof of the asymptotic exactness of the approximation for this case. This is so for the following two reasons. First, Proposition 2 therein assumes the quantity $X(t)$ is precisely Gaussian without investigating the error in this Gaussian approximation. Second, the analysis after Proposition 2 replaces the erfc function by the step function without further investigating the error introduced.

5 RATE OF CONVERGENCE

In this section, we provide two bounds on the rate of convergence in the TTL approxmation under different sets of assumptions.

The following proposition provides a convergence rate of order $(\log C_n/C_n)^{1/4}$. It is stated for the uniform convergence of hit probabilities assuming (R5), the uniform Lipschitz continuity of $\mathcal{G}$. The obvious modification gives the convergence rate for content i assuming uniform Lipschitz continuity of $\mathcal{G}_{n,i}$. Examples of uniformly Lipschitz continuous cdfs include families of distributions that have densities with a common upper bound.

Proposition 5.1. *Under assumptions* (C1), (R5) *and* (P1), *the following holds,*

$$\max_{1\leq i\leq n} \left| H^{\mathrm{LRU}}_{n,i} - H^{\mathrm{TTL}}_{n,i}(T_n) \right| = O\left(\left(\frac{\log C_n}{C_n} \right)^{\frac{1}{4}} \right). \tag{53}$$

Proof. Let M be the Lipschitz constant in (R5). By setting $\xi(\epsilon) = \epsilon/M$ in the proof of Proposition 4.4, we obtain the following,

$$\max_{1\leq i\leq n} \left| H^{\mathrm{LRU}}_{n,i} - H^{\mathrm{TTL}}_{n,i}(T_n) \right| \leq \epsilon + \exp\left\{ -\frac{(\phi x C_n - 1)^2}{4(1+x)C_n + 4} \right\},$$

for $\frac{1}{\phi C_n} \le x \le \frac{\epsilon^2}{M+\epsilon^2}$. For fixed x, the smallest ϵ is $\epsilon = \sqrt{\frac{xM}{1-x}}$. Thus

$$\max_{1\le i\le n} \left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| \le \sqrt{\frac{xM}{1-x}} + \exp\left\{-\frac{(\phi x C_n - 1)^2}{4(1+x)C_n + 4}\right\}.$$

Let $x = \frac{1}{\phi}\sqrt{\frac{\log C_n}{C_n}}$, which satisfies $\frac{1}{\phi C_n} \le x \le x_0$ when C_n is large enough. Then the first term on the r.h.s. of the above inequality is asymptotically equal to

$$\sqrt{\frac{M}{\phi}}\left(\frac{\log C_n}{C_n}\right)^{\frac{1}{4}} = \Theta\left(\left(\frac{\log C_n}{C_n}\right)^{\frac{1}{4}}\right),$$

while the second term is asymptotically equal to

$$\exp\left\{-\frac{1}{4}\log C_n + o(1)\right\} \sim C_n^{-1/4}.$$

It immediately follows that (53) holds. □

The next proposition provides a faster rate of convergence under a different condition, (R6), which says the change in the value of a cdf is bounded by a constant multiple of the relative change in its argument. In fact, we only need (25) to hold with $t = T_n$. Numerical results (see e.g [17]) show that the approximation may converge faster in practice than suggested by (54).

Proposition 5.2. *Under assumptions* (C1), (R6) *and* (P1), *the following holds,*

$$\max_{1\le i\le n} \left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| = O\left(\sqrt{\frac{\log C_n}{C_n}}\right). \tag{54}$$

Proof. Note that the inequality in (25) is invariant under scaling of t, so (R6) implies that (25) holds for $G_{n,i}, \forall n, i$. Replacing the bounds $G_{n,i}(T_n) - G_{n,i}(T_n^+) \le \epsilon$ and $G_{n,i}(T_n) - G_{n,i}(T_n^-) \le \epsilon$ by (25) in the proof of Proposition 4.4, we obtain the following,

$$\max_{1\le i\le n} \left|H_{n,i}^{\mathrm{LRU}} - H_{n,i}^{\mathrm{TTL}}(T_n)\right| \le Bx + \exp\left\{-\frac{(\phi x C_n - 1)^2}{4(1+x)C_n + 4}\right\},$$

for $\frac{1}{\phi C_n} \le x \le \min\{\rho, x_0\}$. Let $x = \frac{1}{\phi}\sqrt{\frac{2\log C_n}{C_n}}$, which falls in the interval $[(\phi C_n)^{-1}, \min\{\rho, x_0\}]$ when $C_n \ge \max\{2, (\min\{\rho, x_0\}\phi)^{-4}\}$. Then the second term on the r.h.s. of the above inequality is asymptotically equal to

$$\exp\left\{-\frac{1}{2}\log C_n + o(1)\right\} \sim C_n^{-1/2}.$$

It immediately follows that (54) holds. □

The following examples show that (R6) holds for a large class of distributions.

Example 5.3. For Poisson request processes, $\mathcal{G} = \{\Psi\}$ with $\Psi(t) = 1 - e^{-t}$. For any $x \ge 0$,

$$0 \le \Psi(t + xt) - \Psi(t) = e^{-t}(1 - e^{-xt}) \le xte^{-t} \le e^{-1}x,$$

where we have used inequalities $e^{-z} \ge 1 - z$ and $ze^{-z} \le e^{-1}$. For $x \in [0, 1]$,

$$0 \le \Psi(t) - \Psi(t - xt) \le \sup_{z\ge 0} e^{-z}(e^{xz} - 1) = (1 - x)^{\frac{1}{x}-1}x \le x.$$

Thus (R6) holds with $B = 1$ and $\rho = 1$.

Example 5.4. Suppose every $G \in \mathcal{G}$ has continuous density on $(0, \infty)$. By the Mean Value Theorem, there exists $\xi_G \in [1, 1+x]$ such that

$$0 \le G(t+xt) - G(t) = G'(\xi_G t)xt \le \xi_G t G'(\xi_G t)x \le \left[\sup_{t>0} tG'(t)\right] x \le B_0 x,$$

where

$$B_0 = \sup_{G \in \mathcal{G}} \sup_{t>0} tG'(t).$$

Similarly, there exists $\zeta_G \in [1-x, 1]$ such that

$$0 \le G(t) - G(t-xt) = G'(\zeta_G t)xt \le \frac{\zeta_G}{1-x} tG'(\zeta_G t)x \le \frac{x}{1-x}\left[\sup_{t>0} tG'(t)\right] \le \frac{B_0}{1-x}x.$$

If $B_0 < \infty$, then (R6) holds with any $\rho \in (0,1)$ and $B = \frac{B_0}{1-\rho}$. When is $B_0 < \infty$ then? Since G has finite mean, $\sup_{t>0} tG'(t) < \infty$. If $\mathcal{G}$ is finite, i.e. the $G_{n,i}$'s are from a finite number of scale families, then $B_0 < \infty$ after taking the supremum over a finite set. In particular, for Poisson request processes, $\mathcal{G} = \{\Psi\}$ with $\Psi(t) = 1 - e^{-t}$, so

$$B_0 = \sup_{t>0} t\Psi'(t) = \sup_{t>0} te^{-t} = e^{-1} < \infty.$$

Thus (R6) holds with any $\rho \in (0,1)$ and $B = e^{-1}(1-\rho)^{-1}$, which is weaker than what we have obtained in Example 5.3.

However, when $\mathcal{G}$ is infinite, i.e., the $G_{n,i}$'s are not from a finite number of scale families, B_0 may still diverge to infinity when we take the supremum over $G \in \mathcal{G}$. An example where we still have finite B_0 is provided by an infinite collection of gamma distributions with shape parameters upper bounded by some $\alpha_{\max} < \infty$. Recall that a gamma distribution G_α with unit mean and shape parameter $\alpha > 0$ has the following density,

$$G'_\alpha(t) = \frac{\alpha^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\alpha t}, \quad t > 0.$$

Hence

$$\sup_{t>0} tG'_\alpha(t) = \frac{\alpha^\alpha}{\Gamma(\alpha)} \sup_{t>0} t^\alpha e^{-\alpha t} = \frac{\alpha^\alpha}{\Gamma(\alpha)} \left(\sup_{t>0} te^{-t}\right)^\alpha = \frac{\alpha^\alpha e^{-\alpha}}{\Gamma(\alpha)},$$

and

$$B_0 = \sup_{\alpha: G_\alpha \in \mathcal{G}} \frac{\alpha^\alpha e^{-\alpha}}{\Gamma(\alpha)} \le \sup_{0<\alpha\le\alpha_{\max}} \frac{\alpha^\alpha e^{-\alpha}}{\Gamma(\alpha)}.$$

Since the function $\alpha^\alpha e^{-\alpha}/\Gamma(\alpha)$ is continuous and has limit 0 as $\alpha \to 0$, we obtain $B_0 < \infty$. Note that as $\alpha \to \infty$,

$$\frac{\alpha^\alpha e^{-\alpha}}{\Gamma(\alpha)} \sim \sqrt{2\pi\alpha} \to \infty,$$

so the boundedness of α is essential.

Corollary 5.5. *Assume $C_n \le \beta_1 n$ for some $\beta_1 \in (0,1)$ and the popularity distribution is Zipf's law in (11). Then (54) holds if $m_\Psi = 1$ and Ψ has a continuous density. In particular, (54) holds if all request processes are Poisson.*

Proof. We check the assumptions of Proposition 5.2. Condition (C1) is assumed. Condition (P1) holds for Zipfian popularity by Example 4.5. By Example 5.4, condition (R6) holds when $m_\Psi = 1$ and Ψ has a continuous density. □

6 EXTENSION OF FAGIN'S RESULT

In this section, we derive expressions for the characteristic time and the aggregate hit probability in the limit as the cache size and the number of contents go to infinity. This extends the results of Fagin [14] for the independence reference model to the more general setting of independent stationary and ergodic content request processes.

We first consider the case where $m_\Psi = 1$ and $p_{n,i} \sim g_n f(z_{n,i})$ uniformly for some continuous function f defined on $(0,1]$ and $z_{n,i} \in [\frac{i-1}{n}, \frac{i}{n}]$, i.e. (R4) and (P2) hold. Recall that $m_\Psi = 1$ implies the cdfs $G_{n,i}$ are all from the same scale family, i.e. $G_{n,i}(t) = \Psi(\lambda_{n,i} t)$ for all n and i.

The following proposition gives the asymptotic expression of T_n, which will be used in the proof of Proposition 6.3 and is also of independent interest. Note that (55) is a generalization of Eq. (2.2) of [14] and Eq. (7) of [17]. We have imposed the inessential condition $f > 0$ a.e. on $[0,1]$, which simplifies the statements and can be easily removed. The proof is found in Section 6.1.

Proposition 6.1. *Under assumptions* (C2), (R4) *and* (P2), *the following holds*

$$T_n \sim \frac{v_0}{g_n \Lambda_n}, \tag{55}$$

where v_0 *the unique real number in* $(0,\infty)$ *that satisfies*

$$\int_0^1 \hat{\Psi}(v_0 f(x))dx = \beta_0. \tag{56}$$

Example 6.2. Consider Zipf's law in (11) with $\alpha \geq 0$. Then $p_{n,i} \sim g_n f(i/n)$ with $f(x) = x^{-\alpha}$ and

$$g_n = \begin{cases} \frac{1-\alpha}{n}, & \text{if } \alpha < 1; \\ \frac{1}{n \log n}, & \text{if } \alpha = 1; \\ \frac{1}{\zeta(\alpha) n^\alpha}, & \text{if } \alpha > 1. \end{cases}$$

It is easy to check that

$$\max_{1 \leq i \leq n} \left| \frac{g_n f(i/n)}{p_{n,i}} - 1 \right| = \left| g_n n^\alpha \sum_{j=1}^n j^{-\alpha} - 1 \right| \to 0$$

as $n \to \infty$, so (P2) holds. If (C2) and (R4) also hold, then T_n satisfies (55). In particular, if $\lambda_{n,i} = i^{-\alpha}$, then $g_n \Lambda_n \sim n^{-\alpha}$ and hence $T_n \sim v_0 n^\alpha$. For Poisson request processes, $\hat{\Psi}(t) = 1 - e^{-t}$ and we recover Eq. (7) of [17].

The following proposition gives the limiting aggregate hit probability, which generalizes Eq. (2.3) of [14]. The proof is found in Section 6.2.

Proposition 6.3. *Assume* (C2), (R4) *and* (P2) *with* $g_n = n^{-1}$. *Then,*

$$H_n^{\text{LRU}} \to \int_0^1 f(x) \Psi(v_0 f(x)) dx, \tag{57}$$

as $n \to \infty$, *where* v_0 *satisfies* (56).

Proposition 6.3 considers a single class of contents in the sense that there is a single f and a single Ψ for all contents. Consider the following generalization to a setting with multiple classes of contents, which may arise from a situation where multiple service providers share a common LRU cache. More precisely, consider J classes of contents, where class j has $b_j n$ contents[8] with $b_j > 0$

[8] We assume $b_j n$ is an integer for ease of presentation, but this can easily relaxed by requiring class j to have a fraction b_j of the contents asymptotically.

and $\sum_{j=1}^{J} b_j = 1$. Instead of labeling contents by a single index i, we label them by a double index so that (j,k) is the k-th content belonging to class j. Correspondingly, we have $\lambda_{n,j,k}$ instead of $\lambda_{n,i}$, and similarly for other quantities. For each class j,

(a) the inter-request distributions are from the same scale family, i.e. $G_{n,j,k}(x) = \Psi_j(\lambda_{n,j,k}x)$ for some continuous cdf Ψ_j with support in $[0,\infty)$ and $m_{\Psi_j} = 1$;

(b) the content popularities $p_{n,j,k} \sim n^{-1} f_j(z_{n,j,k})$ uniformly in k for $z_{n,j,k} \in [\frac{k-1}{b_j n}, \frac{k}{b_j n}]$ and continuous function f_j defined on $(0,1]$ such that $f_j > 0$ a.e. and $\lim_{x\to 0+} f_j(x) \in [0,+\infty]$, i.e.

$$\max_{1\le k\le n_j} \left| \frac{f_j(z_{n,j,k})}{np_{n,j,k}} - 1 \right| \to 0, \quad \text{as } n \to \infty; \tag{58}$$

Note that (a) implies (R3) and (b) is the precise statement of (P3). We have the following generalization of Proposition 6.3. The proof is found in Appendix B.

Proposition 6.4. *Assume* (C2), *and conditions* (a) *and* (b) *above. Then*

$$H_n^{\mathrm{LRU}} \to \sum_{j=1}^{J} b_j \int_0^1 f_j(x)\Psi_j(v_0 f_j(x))dx, \tag{59}$$

as $n \to \infty$, where v_0 is the unique real number in $(0,\infty)$ that satisfies

$$\sum_{j=1}^{J} b_j \int_0^1 \hat{\Psi}_j(v_0 f_j(x))dx = \beta_0. \tag{60}$$

6.1 Proof of Proposition 6.1

We need the following lemmas.

Lemma 6.5. *The function*

$$\beta(v) := \int_0^1 \hat{\Psi}(vf(x))dx$$

has the following properties,

(i) $\beta(0) = 0$, $\lim_{v\to\infty} \beta(v) = 1$;
(ii) β *is continuous;*
(iii) β *is increasing in* v;
(iv) β *is strictly increasing at all* v *such that* $\beta(v) < 1$.

Proof. By (3), $\hat{\Psi}(0) = 0$, which implies in turn implies that $\beta(0) = 0$. Since $\lim_{t\to\infty} \hat{\Psi}(t) = 1$, by the Bounded Convergence Theorem,

$$\lim_{v\to\infty} \beta(v) = \int_0^1 \lim_{v\to\infty} \hat{\Psi}(vf(x))dx = 1.$$

This proves (i). (ii) follows from the continuity of $\hat{\Psi}$ and the Bounded Convergence Theorem.

Let $v_1 > v_2$. Since $f(x) \ge 0$, it follows that $\hat{\Psi}(v_1 f(x)) \ge \hat{\Psi}(v_2 f(x))$ and hence $\beta(v_1) \ge \beta(v_2)$. This proves (iii).

If $\beta(v_1) = \beta(v_2)$, continuity of $\hat{\Psi}(vf(x))$ implies $\hat{\Psi}(v_1 f(x)) = \hat{\Psi}(v_2 f(x))$ for all x. If $f(x) > 0$, then $v_1 f(x) > v_2 f(x)$, and (3) implies $\bar{\Psi}(v_2 f(x)) = 0$, which, by monotonicity of $\bar{\Psi}$, implies $\bar{\Psi}(t) = 0$ for all $t \ge v_2 f(x)$. Thus

$$1 - \hat{\Psi}(v_2 f(x)) = \int_{v_2 f(x)}^{\infty} \bar{\Psi}(t)dt = 0.$$

It follows that $\hat{\Psi}(\nu_2 f(x)) = \mathbb{1}_{\{f(x)>0\}}$ for all $x \in (0,1]$. Since $\hat{\Psi}(\nu_2 f(x))$ is continuous in x and f is not identically zero, it follows that $\hat{\Psi}(\nu_2 f(x)) = 1$ and hence $\beta(\nu_2) = 1$. Thus $\beta(\nu_1) > \beta(\nu_2)$ if $\beta(\nu_2) < 1$, which completes the proof of (iv). □

Now we prove Proposition 6.1.

Proof of Proposition 6.1. Recall $m_\Psi = 1$ implies $G_{n,i}(x) = \Psi(\lambda_{n,i}x)$ and $\hat{G}_{n,i}(x) = \hat{\Psi}(\lambda_{n,i}x)$. We obtain from (36) and (9),

$$\frac{C_n}{n} = \frac{1}{n}K_n(T_n) = \frac{1}{n}\sum_{i=1}^{n}\hat{G}_{n,i}(T_n) = \frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}(\lambda_{n,i}T_n) = \frac{1}{n}\sum_{i=1}^{n}\hat{\Psi}(p_{n,i}\Lambda_n T_n).$$

Given any $\epsilon > 0$, (27) yields that for sufficiently large n and $i = 1, \ldots, n$,

$$(1-\epsilon)g_n f(z_{n,i}) \le p_{n,i} \le (1+\epsilon)g_n f(z_{n,i}). \tag{61}$$

Let $\nu_1 = \limsup_{n\to\infty} g_n\Lambda_n T_n$. Let $\{n_\ell : \ell \ge 1\}$ be the indices of a subsequence that converges to ν_1, i.e. $\nu_1 = \lim_{\ell\to\infty} g_{n_\ell}\Lambda_{n_\ell}T_{n_\ell}$. First assume $\nu_1 < \infty$. For sufficiently large ℓ,

$$(1-\epsilon)(\nu_1-\epsilon)f(z_{n_\ell,i}) \le p_{n_\ell,i}\Lambda_{n_\ell}T_{n_\ell} \le (1+\epsilon)(\nu_1+\epsilon)f(z_{n_\ell,i}).$$

Since $\hat{\Psi}$ is non-decreasing, for sufficiently large ℓ,

$$\begin{aligned}\frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\hat{\Psi}\left((1-\epsilon)(\nu_1-\epsilon)f(z_{n_\ell,i})\right) \le \frac{C_{n_\ell}}{n_\ell} &= \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\hat{\Psi}(p_{n_\ell,i}\Lambda_{n_\ell}T_{n_\ell}) \\ &\le \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}\hat{\Psi}\left((1+\epsilon)(\nu_1+\epsilon)f(z_{n_\ell,i})\right).\end{aligned}$$

Letting $\ell \to \infty$ and using the definition of the Riemann integral, we obtain

$$\int_0^1 \hat{\Psi}((1-\epsilon)(\nu_1-\epsilon)f(x))dx \le \lim_{k\to\infty}\frac{C_{n_\ell}}{n_\ell} = \beta_0 \le \int_0^1 \hat{\Psi}((1+\epsilon)(\nu_1+\epsilon)f(x))dx.$$

Since $\hat{\Psi}$ is continuous, letting $\epsilon \to 0$ and using the Bounded Convergence Theorem, we obtain

$$\beta_0 = \int_0^1 \hat{\Psi}(\nu_1 f(x))dx = \beta(\nu_1).$$

If $\nu_1 = +\infty$, repeating the above argument shows that

$$\beta_0 \ge \beta(\nu)$$

for any ν, which would imply $\beta_0 \ge \lim_{\nu\to\infty}\beta(\nu) = 1$ by Lemma 6.5, a contradiction. Therefore, ν_1 is finite and satisfies $\beta(\nu_1) = \beta_0$. The same argument shows that $\nu_2 = \liminf_{n\to\infty} g_n\Lambda_n T_n$ satisfies $\beta_0 = \beta(\nu_2)$. By Lemma 6.5, $\nu_1 = \nu_2 = \nu_0$, where $\nu_0 \in (0,\infty)$ is the unique root of $\beta(\nu) = \beta_0$. It follows that (55) holds. □

6.2 Proof of Proposition 6.3

We will invoke Corollary 4.6 to show convergence. Assumption (R2) holds by Lemma A.1. Since $m_\Psi = 1$ by (R4), (C2) implies (C1) for any $\beta_1 \in (\beta_0, 1)$.

Now we show that (P1) holds. Let $A_\ell = \{x \in [0,1] : f(x) \ge 1/\ell\}$. Since $f > 0$ a.e., $\lim_{\ell\to\infty} Leb(A_\ell) = Leb\{x \in [0,1] : f(x) > 0\} = 1$, where Leb is the Lebesgue measure on $[0,1]$. Thus there exists an ℓ_0 such that $Leb(A^c_{\ell_0}) \le (1-\kappa_1\beta_0)/4$. Let $I = [(1-\kappa_1\beta_0)/4, 1]$ and

$I_{n,i} = [\frac{i-1}{n}, \frac{i}{n}]$. Since f is continuous, it is uniformly continuous on I by the Heine-Cantor Theorem. For all sufficiently large n, $|f(x) - f(z_{n,i})| \le \frac{1}{2\ell_0}$ if $x \in I_{n,i} \cap I$. Therefore, for all sufficiently large n,

$$\begin{aligned}\frac{1}{n} f(z_{n,i}) = \int_{I_{n,i}} f(z_{n,i}) dx &\ge \int_{I_{n,i}\cap I} \left[f(x) - \frac{1}{2\ell_0} \right] dx \\ &\ge \int_{I_{n,i}\cap I\cap A_{\ell_0}} \left(\frac{1}{\ell_0} - \frac{1}{2\ell_0} \right) dx = \frac{1}{2\ell_0} Leb(I_{n,i} \cap I \cap A_{\ell_0}).\end{aligned}$$

Summing over i, we obtain

$$\begin{aligned}\bar{P}_n(\lceil \kappa_1 C_n \rceil) &\sim \sum_{i=\lceil \kappa_1 C_n \rceil + 1}^{n} \frac{1}{n} f(z_{n,\sigma_i}) \\ &\ge \frac{1}{2\ell_0} \sum_{i=\lceil \kappa_1 C_n \rceil + 1}^{n} Leb(I_{n,\sigma_i} \cap I \cap A_{\ell_0}) \\ &= \frac{1}{2\ell_0} Leb\left(\left(\bigcup_{i=\lceil \kappa_1 C_n \rceil + 1}^{n} I_{n,\sigma_i} \right) \cap I \cap A_{\ell_0} \right) \\ &\ge \frac{1}{2\ell_0} \left(Leb\left(\bigcup_{i=\lceil \kappa_1 C_n \rceil + 1}^{n} I_{n,\sigma_i} \right) - Leb(I^c) - Leb(A_{\ell_0}^c) \right) \\ &= \frac{1}{2\ell_0} \left(\sum_{i=\lceil \kappa_1 C_n \rceil + 1}^{n} Leb(I_{n,\sigma_i}) - Leb(I^c) - Leb(A_{\ell_0}^c) \right) \\ &\ge \frac{1}{2\ell_0} \left(\frac{n - \lceil \kappa_1 C_n \rceil}{n} - \frac{1}{4}(1 - \kappa_1 \beta_0) - \frac{1}{4}(1 - \kappa_1 \beta_0) \right) \\ &= \frac{1}{4\ell_0}(1 - \kappa_1 \beta_0) > 0. \end{aligned} \tag{62}$$

We conclude that (26) holds for $0 < \gamma < \frac{1}{4\ell_0}(1 - \kappa_1 \beta_0)$.

Therefore, (32) holds by Corollary 4.6. Then (57) follows from (32) and the following lemma.

Lemma 6.6. *Under the assumptions of Proposition 6.3,*

$$H_n^{\mathrm{TTL}}(T_n) \to \int_0^1 f(x) \Psi(\nu_0 f(x)) dx, \quad \text{as } n \to \infty. \tag{63}$$

Proof. Recall that $G_{n,i}(x) = \Psi(\lambda_{n,i} x)$. We obtain from (20), (21) and (34),

$$H_n^{\mathrm{TTL}}(T_n) = \sum_{i=1}^{n} p_{n,i} \Psi(\lambda_{n,i} T_n) = \sum_{i=1}^{n} p_{n,i} \Psi(p_{n,i} \Lambda_n T_n).$$

From (55) and (61) the following inequalities hold, for any $\epsilon > 0$ and n large enough,

$$(1 - \epsilon)(\nu_0 - \epsilon) f(z_{n,i}) \le p_{n,i} \Lambda_n T_n \le (1 + \epsilon)(\nu_0 + \epsilon) f(z_{n,i}).$$

The monotonicity of Ψ then yields

$$\begin{aligned}\frac{1-\epsilon}{n} \sum_{i=1}^{n} f(z_{n,i}) \Psi\left((1-\epsilon)(\nu_0 - \epsilon) f(z_{n,i}) \right) &\le H_n^{\mathrm{TTL}}(T_n) \\ &\le \frac{1+\epsilon}{n} \sum_{i=1}^{n} f(z_{n,i}) \Psi\left((1+\epsilon)(\nu_0 + \epsilon) f(z_{n,i}) \right).\end{aligned} \tag{64}$$

Letting $n \to \infty$ and using the definition of the Riemann integral, we obtain

$$\liminf_{n\to\infty} H_n^{\mathrm{TTL}}(T_n) \geq (1-\epsilon)\int_0^1 f(x)\Psi((1-\epsilon)(v_0-\epsilon)f(x)dx, \tag{65}$$

and

$$\limsup_{n\to\infty} H_n^{\mathrm{TTL}}(T_n) \leq (1+\epsilon)\int_0^1 f(x)\Psi((1+\epsilon)(v_0+\epsilon)f(x))dx. \tag{66}$$

The existence of the integrals comes from the fact that $0 \leq \Psi \leq 1$ and the integrability of f over $[0,1]$, which follows from the first inequality in (61) by the following,

$$1 = \sum_{i=1}^{n} p_{n,i} \geq (1-\epsilon)\frac{1}{n}\sum_{i=1}^{n} f(z_{n,i}) \to (1-\epsilon)\int_0^1 f(x)dx.$$

Since Ψ is continuous and $\int_0^1 f(x)dx < \infty$, letting $\epsilon \to 0$ in (65) and (66) yields (63) by the Dominated Convergence Theorem. □

7 CONCLUSIONS

In this paper, we developed an approximation for the aggregate and individual content hit probability of an LRU cache based on a transformation to the TTL cache for the case that content requests are described by independent stationary and ergodic processes. This approximation extends one first proposed and studied by Fagin [14] for the independent reference model and provides the theoretical basis for approximations introduced in [18] for content requests described by independent renewal processes. We showed that the approximations become exact in the limit as the cache size and the number of contents go to infinity. Last, we established the rate of convergence for the approximation as number of contents increases.

Future directions include investigation for tighter bounds on the convergence rate and extension of these results to other cache policies such as FIFO and random and to networks of caches perhaps using ideas from [4, 9, 28]. In addition, it is desirable to relax independence between different content request streams.

REFERENCES

[1] *Private communication between P. Jelenkovic and the authors.*

[2] T. M. Apostol. 1976. *Introduction to Analytic Number Theory.* Springer-Verlag.

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

[4] D. S. Berger, P. Gland, S. Singla, and F. Ciucu. 2014. Exact analysis of TTL cache networks. *Performance Evaluation* 79 (2014), 2–23.

[5] G. Bianchi, A. Detti, A. Caponi, and N. Blefari-Melazzi. 2013. Check before storing: What is the performance price of content integrity verification in LRU caching? *ACM SIGCOMM Computer Communication Review* 43, 3 (2013), 59–67.

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

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

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

[9] N. Choungmo Fofack, M. Dehghan, D. Towsley, M. Badov, and D. Goeckel. 2014. On the Performance of General Cache Networks. In *Proceedings ValueTools 2014.* Bratislavia, Slovakia.

[10] N. Choungmo Fofack, P. Nain, G. Neglia, and D. Towsley. 2014. Performance Evaluation of Hierarchical TTL-based Cache Networks. *Computer Networks* 65 (June 2014), 212–231.

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

[12] A. Dan and D. Towsley. 1990. An approximate analysis of the LRU and FIFO buffer replacement schemes. In *Proc. ACM SIGMETRICS.* Boulder, CO, USA, 143–152.

[13] M. Dehghan, B. Jiang, A. Dabirmoghaddam, and D. Towsley. 2015. On the analysis of caches with pending interest tables. In *Proceedings of the 2nd International Conference on Information-centric Networking*. ACM, 69–78.
[14] R. Fagin. 1977. Asymptotic Miss Ratios over Independent References. *J. Comput. System Sci.* 14, 2 (1977), 222–250.
[15] A. Ferragut, I. Rodríguez, and F. Paganini. 2016. Optimizing TTL caches under heavy-tailed demands. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*. 101–112.
[16] P. Flajolet, D. Gardy, and L. Thimonier. 1992. Birthday paradox, coupon collector, caching algorithms and self-organizing search. *Discrete Applied Mathematics* 39 (1992), 207–229.
[17] C. Fricker, P. Robert, and J. Roberts. 2012. A Versatile and Accurate Approximation for LRU Cache Performance. In *Proceedings of the 24th International Teletraffic Congress (ITC 24)*. Kraków, Poland.
[18] M. Garetto, E. Leonardi, and V. Martina. 2016. 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).
[19] N. Gast and B. Van Houdt. 2017. TTL approximations of the cache replacement algorithms LRU(m) and h-LRU. *Performance Evaluation* (2017). http://dx.doi.org/10.1016/j.peva.2017.09.002.
[20] R. Hirade and T. Osogami. 2010. Analysis of page replacement policies in the fluid limit. *Operations research* 58, 4-part-1 (2010), 971–984.
[21] P. Jelenkovic and A. Radovanović. 2004. Least-recently used caching with dependent requests. *Theoretical Computer Science* 326 (2004), 293–327.
[22] P. Jelenkovic, A. Radovanović, and M. Squillante. 2006. Critical sizing of LRU caches with dependent requests. *J. of Applied Probability* 43, 4 (2006), 1013–1027.
[23] P. R. Jelenković. 1999. Asymptotic approximation of the move-to-front search cost distribution and least-recently used caching fault probabilities. *Annals of Applied Probability* (1999), 430–464.
[24] W. F. King. 1972. Analysis of demand paging algorithm. *Information Processing* 71 (1972), 485–490.
[25] E. Leonardi and G. L. Torrisi. 2017. Modeling Least Recently Used caches with Shot Noise request processes. *SIAM J. Appl. Math.* 77, 2 (2017), 361–383.
[26] M. Loève. 1977. *Probability Theory I* (4th ed.). Graduate Texts in Mathematics, Vol. 45. Springer-Verlag, New York.
[27] T. Osogami. 2010. A fluid limit for a cache algorithm with general request processes. *Advances in Applied Probability* 42, 3 (2010), 816–833.
[28] E. J. Rosensweig, J. Kurose, and D. Towsley. 2010. Approximate Models for General Cache Networks. In *Proceedings of Infocom 2010*. San Diego, CA, USA, 1100–1108.

A EQUICONTINUITY

Lemma A.1. *A finite family of continuous cdfs is equicontinuous, so that* (R4) $\Longrightarrow$ (R3) $\Longrightarrow$ (R2).

Proof. Let the family of cdfs be $\mathcal{F} = \{F_1, \ldots, F_J\}$. Fix ϵ. There exists a $L_j \in (0, \infty)$ such that

$$F_j(-L_j) < \epsilon \quad \text{and} \quad 1 - F_j(L_j) < \epsilon. \tag{67}$$

Let $L = \max_{1 \le j \le J} L_j \in (0, \infty)$. Being continuous, F_j is uniformly continuous on $[-2L, 2L]$ by the Heine-Cantor Theorem. Thus there exists a $\delta_j \in (0, L)$ such that

$$|F_j(x_1) - F_j(x_2)| < \epsilon, \tag{68}$$

for $x_1, x_2 \in [-2L, 2L]$ such that $|x_1 - x_2| < \delta_j$.

Let $\delta = \min_{1 \le j \le J} \delta_j \in (0, L)$. Consider any $x_1 > x_2$ with $|x_1 - x_2| < \delta$. There are three cases.

(i) If $x_1, x_2 \in [-2L, 2L]$, then (68) holds for all j.
(ii) If $x_1 > 2L$, then $x_2 > L$, since $|x_1 - x_2| < \delta < L$. Thus $|F_j(x_1) - F_j(x_2)| = F_j(x_1) - F_j(x_2) \le 1 - F_j(L_j) < \epsilon$ by (67), and this holds for all j.
(iii) If $x_2 < -2L$, then $x_1 < -L$, since $|x_1 - x_2| < \delta < L$. Thus $|F_j(x_1) - F_j(x_2)| =\le F_j(-L_j) < \epsilon$ by (67), and this holds for all j.

Therefore, $\mathcal{F}$ is equicontinuous. □

B PROOF OF PROPOSITION 6.4

The proof parallels those of Proposition 6.1 and 6.3 except for the last step. The following lemma generalizes Lemma 6.5.

Lemma B.1. *The function*

$$\beta_J(v) := \sum_{j=1}^{J} b_j \int_0^1 \hat{\Psi}_j(v f_j(x))dx$$

has the following properties,

(i) $\beta_J(0) = 0$, $\lim_{v\to\infty} \beta_J(v) = 1$;
(ii) β_J *is continuous;*
(iii) β_J *is increasing in* v;
(iv) β_J *is strictly increasing at all* v *such that* $\beta_J(v) < 1$.

Proof. By (8), $\hat{\Psi}_j(0) = 0$, which implies $\beta_J(0) = 0$. Since $\lim_{x\to\infty} \hat{\Psi}_j(x) = 1$, by the Bounded Convergence Theorem,

$$\lim_{v\to\infty} \beta_J(v) = \sum_{j=1}^{J} b_j \int_0^1 \lim_{v\to\infty} \hat{\Psi}_j(v f_j(x))dx = 1.$$

This proves (i). (ii) follows from the continuity of $\hat{\Psi}_j$ and the Bounded Convergence Theorem.

Let $v_1 > v_2$. Since $f_j(x) \geq 0$, it follows that $\hat{\Psi}_j(v_1 f_j(x)) \geq \hat{\Psi}_j(v_2 f_j(x))$ and hence $\beta_J(v_1) \geq \beta_J(v_2)$. This proves (iii).

If $\beta_J(v_1) = \beta_J(v_2)$, then continuity of $\hat{\Psi}_j(v f_j(x))$ implies that $\hat{\Psi}_j(v_1 f_j(x)) = \hat{\Psi}_j(v_2 f_j(x))$ for all x. If $f_j(x) > 0$, then $v_1 f_j(x) > v_2 f_j(x)$, and (3) implies $\bar{\Psi}_j(v_2 f_j(x)) = 0$, which, by monotonicity of $\bar{\Psi}_j$, implies $\bar{\Psi}_j(t) = 0$ for all $t \geq v_2 f_j(x)$. Thus

$$1 - \hat{\Psi}_j(v_2 f_j(x)) = \int_{v_2 f_j(x)}^{\infty} \bar{\Psi}_j(t)dt = 0.$$

It follows that $\hat{\Psi}_j(v_2 f_j(x)) = \mathbb{1}_{\{f_j(x)>0\}}$ for all $x \in (0, 1]$. Since $\hat{\Psi}_j(v_2 f_j(x))$ is continuous in x and f_j is not identically zero, it follows that $\hat{\Psi}_j(v_2 f_j(x)) = 1$. Since this is true for all j, it follows that $\beta_J(v_2) = 1$. Thus $\beta_J(v_1) > \beta_J(v_2)$ if $\beta_J(v_2) < 1$, which completes the proof of (iv). □

The following proposition generalizes Proposition 6.1.

Proposition B.2. *Under the assumptions in Proposition 6.4 but with the condition (b) that* $p_{n,j,k} \sim n^{-1} f_j(z_{n,j,k})$ *generalized to* $p_{n,j,k} \sim g_n f_j(z_{n,j,k})$, *we have*

$$T_n \sim \frac{v_0}{g_n \Lambda_n}, \tag{69}$$

where v_0 *satisfies (60).*

Proof. Recall $G_{n,j,k}(x) = \Psi_j(\lambda_{n,j,k} x)$ implies $\hat{G}_{n,j,k}(x) = \hat{\Psi}_j(\lambda_{n,j,k} x)$. We obtain from (36) and (9),

$$\frac{C_n}{n} = \frac{1}{n} K_n(T_n) = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{b_j n} \hat{G}_{n,j,k}(T_n) = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{b_j n} \hat{\Psi}_j(\lambda_{n,j,k} T_n) = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{b_j n} \hat{\Psi}_j(p_{n,j,k} \Lambda_n T_n).$$

Given any $\epsilon > 0$, (58) yields that for sufficiently large n and $j = 1, \ldots, J$, $k = 1, \ldots, b_j n$,

$$(1 - \epsilon) g_n f_j(z_{n,j,k}) \leq p_{n,j,k} \leq (1 + \epsilon) g_n f_j(z_{n,j,k}). \tag{70}$$

Let $v_1 = \limsup_{n\to\infty} g_n \Lambda_n T_n$. Let $\{n_\ell : \ell \geq 1\}$ be the indices of a subsequence that converges to v_1, i.e. $v_1 = \lim_{\ell\to\infty} g_{n_\ell} \Lambda_{n_\ell} T_{n_\ell}$. First assume $v_1 < \infty$ for all j. For sufficiently large ℓ,

$$(1 - \epsilon)(v_1 - \epsilon) f_j(z_{n_\ell,i}) \leq p_{n_\ell,j,k} \Lambda_{n_\ell} T_{n_\ell} \leq (1 + \epsilon)(v_1 + \epsilon) f_j(z_{n_\ell,i}).$$

Since $\hat{\Psi}_j$ is non-decreasing, for sufficiently large ℓ,

$$\frac{1}{n_\ell}\sum_{j=1}^{J}\sum_{k=1}^{b_j n_\ell}\hat{\Psi}_j\left((1-\epsilon)(v_1-\epsilon)f_j(z_{n_\ell,i})\right) \le \frac{C_{n_\ell}}{n_\ell} = \frac{1}{n_\ell}\sum_{j=1}^{J}\sum_{k=1}^{b_j n_\ell}\hat{\Psi}_j(p_{n_\ell,i}\Lambda_{n_\ell}T_{n_\ell})$$
$$\le \frac{1}{n_\ell}\sum_{j=1}^{J}\sum_{k=1}^{b_j n_\ell}\hat{\Psi}_j\left((1+\epsilon)(v_1+\epsilon)f_j(z_{n_\ell,i})\right).$$

Letting $\ell \to \infty$ and using the definition of the Riemann integral, we obtain

$$\sum_{j=1}^{J} b_j \int_0^1 \hat{\Psi}_j((1-\epsilon)(v_1-\epsilon)f_j(x))dx \le \lim_{\ell\to\infty}\frac{C_{n_\ell}}{n_\ell} = \beta_0 \le \sum_{j=1}^{J} b_j \int_0^1 \hat{\Psi}_j((1+\epsilon)(v_1+\epsilon)f_j(x))dx.$$

Since $\hat{\Psi}_j$ is continuous, letting $\epsilon \to 0$ and using the Bounded Convergence Theorem, we obtain

$$\beta_0 = \sum_{j=1}^{J} b_j \int_0^1 \hat{\Psi}_j(v_1 f_j(x))dx = \beta_J(v_1).$$

If $v_1 = +\infty$, repeating the above argument shows that

$$\beta_0 \ge \beta_J(v)$$

for any v, which would imply $\beta_0 \ge \lim_{v\to\infty}\beta_J(v) = 1$ by Lemma B.1, a contradiction. Therefore, v_1 is finite and satisfies $\beta_J(v_1) = \beta_0$. The same argument shows that $v_2 = \liminf_{n\to\infty} g_n\Lambda_n T_n$ satisfies $\beta_0 = \beta_J(v_2)$. By Lemma B.1, $v_1 = v_2 = v_0$, where $v_0 \in (0,\infty)$ is the unique root of $\beta_J(v) = \beta_0$. It follows that (69) holds. □

The following lemma generalizes Lemma 6.6.

Lemma B.3. *Under the assumptions of Proposition 6.4,*

$$H_n^{\mathrm{TTL}}(T_n) \to \sum_{j=1}^{J} b_j \int_0^1 f_j(x)\Psi_j(v_0 f_j(x))dx. \tag{71}$$

Proof. Recall that $G_{n,j,k}(x) = \Psi_j(\lambda_{n,j,k}x)$. We obtain from (20), (21) and (34),

$$H_n^{\mathrm{TTL}}(T_n) = \sum_{j=1}^{J}\sum_{i=1}^{b_j n} p_{n,j,k}\Psi_j(\lambda_{n,j,k}T_n) = \sum_{j=1}^{J}\sum_{i=1}^{b_j n} p_{n,j,k}\Psi_j(p_{n,j,k}\Lambda_n T_n).$$

Given any $\epsilon > 0$, for all sufficiently large n, (70) and the following hold,

$$(1-\epsilon)(v_0-\epsilon)f_j(z_{n,j,k}) \le p_{n,j,k}\Lambda_n T_n \le (1+\epsilon)(v_0+\epsilon)f_j(z_{n,j,k}). \tag{72}$$

The monotonicity of Ψ_j then yields

$$\frac{1-\epsilon}{n}\sum_{j=1}^{J}\sum_{k=1}^{b_j n} f_j(z_{n,j,k})\Psi_j\left((1-\epsilon)(v_0-\epsilon)f_j(z_{n,j,k})\right) \le H_n^{\mathrm{TTL}}(T_n)$$
$$\le \frac{1+\epsilon}{n}\sum_{j=1}^{J}\sum_{k=1}^{b_j n} f_j(z_{n,j,k})\Psi_j\left((1+\epsilon)(v_0+\epsilon)f_j(z_{n,j,k})\right).$$

Letting $n \to \infty$ and using the definition of the Riemann integral, we find

$$\liminf_{n\to\infty} H_n^{\mathrm{TTL}}(T_n) \geq (1-\epsilon)\sum_{j=1}^{J} b_j \int_0^1 f_j(x)\Psi_j((1-\epsilon)(\nu_0-\epsilon)f_j(x)dx, \tag{73}$$

and

$$\limsup_{n\to\infty} H_n^{\mathrm{TTL}}(T_n) \leq (1+\epsilon)\sum_{j=1}^{J} b_j \int_0^1 f_j(x)\Psi_j((1+\epsilon)(\nu_0+\epsilon)f_j(x))dx. \tag{74}$$

The existence of the integrals comes from the fact that $0 \leq \Psi_j \leq 1$ and the integrability of f_j over $[0,1]$, which follows from the first inequality in (70) by the following,

$$1 = \sum_{j=1}^{J}\sum_{k=1}^{b_j n} p_{n,j,k} \geq (1-\epsilon)\frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{b_j n} f_j(z_{n,j,k}) \to (1-\epsilon)\sum_{j=1}^{J} b_j \int_0^1 f_j(x)dx. \tag{75}$$

Since Ψ_j is continuous and $\int_0^1 f_j(x)dx < \infty$, letting $\epsilon \to 0$ in (73) and (74) yields (71) by the Dominated Convergence Theorem. □

Proof of Proposition 6.4. Thanks to Lemma B.3 and the value of ν_0 given in Proposition B.2 that satisfies (60), we only need to show the convergence of H_n^{LRU} to $H_n^{\mathrm{TTL}}(T_n)$ as $n \to \infty$. For that, we invoke Corollary 4.6. We use Remark 2 and show that (39) holds under the conditions of Proposition 6.4. Repeating the proof of (71), we obtain

$$H_n^{\mathrm{TTL}}((1+x)T_n) \to \sum_{j=1}^{J} b_j \int_0^1 f_j(y)\Psi_j((1+x)\nu_0 f_j(y))dy, \tag{76}$$

as $n \to \infty$. Fix $\epsilon > 0$. Summing (72) over j and k and letting $g_n = 1/n$ yields, for n large enough,

$$(1-\epsilon)\frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{b_j n} f(z_{n,i,k}) \leq \sum_{j=1}^{J}\sum_{k=1}^{b_j n} p_{n,j,k} = 1 \leq (1+\epsilon)\frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{b_j n} f(z_{n,i,k}).$$

Letting $n \to \infty$ we obtain, by the definition of the Riemann integral,

$$(1-\epsilon)\sum_{j=1}^{J} b_j \int_0^1 f_j(y)dy \leq 1 \leq (1+\epsilon)\sum_{j=1}^{J} b_j \int_0^1 f_j(y)dy,$$

from which we conclude that

$$1 = \sum_{j=1}^{J} b_j \int_0^1 f_j(y)dy. \tag{77}$$

Since $\mu_n(T) = 1 - H_n^{\mathrm{TTL}}(T)$, subtracting (76) from (77) yields

$$\mu_n((1+x)T_n) \to \sum_{j=1}^{J} b_j \int_0^1 f_j(y)\bar{\Psi}_j((1+x)\nu_0 f_j(y))dy := \mu(x).$$

Note that μ is continuous by the continuity of f_j, Ψ_j and the Dominated Convergence Theorem. If $\mu(0) > 0$, then there exists $x_0 > 0$ such that $\mu(x_0) \geq \mu(0)/2 > 0$. Thus for sufficiently large n and $T \leq (1+x_0)T_n$,

$$\mu_n(T) \geq \mu_n((1+x_0)T_n) \geq \mu(x_0)/2 \geq \mu(0)/4 > 0.$$

Since $C_n \leq \Lambda_n T_n$ by (29) with $n_2 = 0$, the above inequality yields (39) with $\phi = \mu(0)/4$, provided that $\mu(0) > 0$.

Now we show that $\mu(0) > 0$. Suppose $\mu(0) = 0$. Then

$$\int_0^1 f_j(y)\bar{\Psi}_j(v_0 f_j(y))dy = 0$$

for each j. Since $f_j(y) > 0$ a.e. on $(0, 1]$ and $\bar{\Psi}_j(\cdot) \geq 0$ for each j, it follows that $\bar{\Psi}_j(v_0 f_j(y)) = 0$ a.e. on $(0, 1]$ for each j. Hence, by (3) with $m_{\Psi_j} = 1$ for each j,

$$\beta_J(v_0) = \sum_{j=1}^{J} b_j \int_0^1 \hat{\Psi}_j(v_0 f_j(x))dx = \sum_{j=1}^{J} b_j \int_0^1 \int_0^{v_0 f_j(x)} \bar{\Psi}(y)dydx = 0,$$

which contradicts the result obtained in Proposition B.2 that $\beta_J(v_0) = \beta_0 \in (0, 1)$. Therefore, $\mu(0) > 0$, which completes the proof. □

Remark 4. A proof similar to that of Proposition 6.3 can be done if we let $\Psi = \max_{1\leq j\leq J} \Psi_j$ and restrict the range of β_0 to $\beta_0 < m_\Psi = \int_0^\infty \min_{1\leq j\leq J} \bar{\Psi}_j(x)dx$ (see Section 2.1), a quantity that is in general strictly less than one. This restriction on β_0 is a consequence of condition (C1). It is also worth noting that, since Proposition 6.4 reduces to Proposition 6.3 when $J = 1$, the proof of Proposition 6.4 provides an alternative proof of Proposition 6.3.